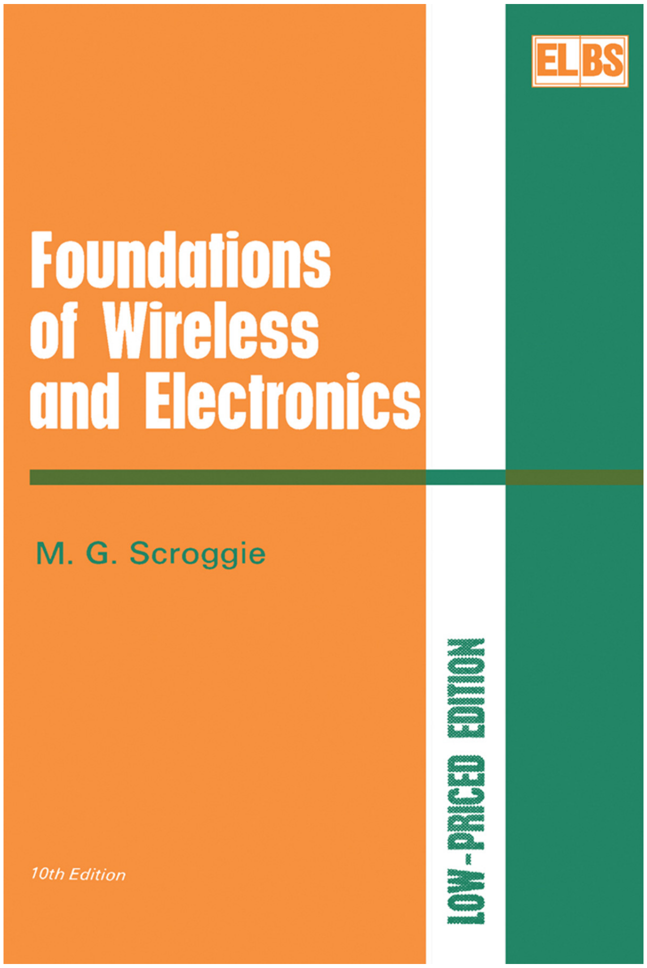

ELBS
Foundations of Wireless and Electronics
M. G. Scroggie
10th Edition
LOW-PRICED EDITION

Foundations of Wireless and Electronics, 10th Edition

M. G. Scroggie, S. W. Amos

Newnes

Table of Contents

Front Cover
Foundations of Wireless and Electronics
Copyright Page
Table of Contents (1/2)
Table of Contents (2/2)
Preface to Ninth Edition
Preface to Tenth Edition
1 General View of a System
  1.1 What Wireless Does
  1.2 Nature of Sound Waves
  1.3 Characteristics of Sound Waves
  1.4 Frequency
  1.5 Wavelength
  1.6 The Sender
  1.7 The Receiver
  1.8 Electrical Communication by Wire
  1.9 Electric Waves
  1.10 Why High Frequencies are Necessary
  1.11 Radio Telegraphy
  1.12 Tuning
  1.13 Radio Telephony
  1.14 Recapitulation
Chapter 2. Electricity and Circuits
  2.1 Electrons
  2.2 Electric Charges and Currents
  2.3 Conductors and Insulators
  2.4 Electromotive Force
  2.5 Electrical Units
  2.6 Ohm's Law
  2.7 Larger and Smaller Units
  2.8 Circuit Diagrams
  2.9 Resistances in Series and in Parallel
  2.10 Series-Parallel Combinations
  2.11 Resistance Analysed
  2.12 Conductance
  2.13 Kirchhoff's Laws

- 2.14 P.D. and E.M.F.
- 2.15 Electrical Effects
- 2.16 Instruments for Measuring Electricity
- 2.17 Electrical Power
- 2.18 A Broader View of Resistance

Chapter 3. Capacitance

- 3.1 Charging Currents
- 3.2 Capacitance: What It Is
- 3.3 Capacitance Analysed
- 3.4 Capacitors
- 3.5 Charge and Discharge of a Capacitor
- 3.6 Where the Power Goes
- 3.7 Recapitulation
- 3.8 Displacement Currents

Chapter 4. Inductance

- 4.1 Magnets and Electromagnets
- 4.2 Interacting Magnetic Fields
- 4.3 Induction
- 4.4 Self-Inductance
- 4.5 Lenz's Law
- 4.6 Inductance Analysed
- 4.7 Practical Considerations
- 4.8 Growth of Current in an Inductive Circuit
- 4.9 Power During Growth
- 4.10 More Comparison and Contrast
- 4.11 Mutual Inductance

Chapter 5. Alternating Currents

- 5.1 Frequencies of Alternating Current
- 5.2 The Sine Wave
- 5.3 Circuit with Resistance Only
- 5.4 R.M.S. Values
- 5.5 A.C. Meters
- 5.6 Phase
- 5.7 Phasor Diagrams
- 5.8 Adding Alternating Voltages
- 5.9 Direction Signs
- 5.10 Subscript Notation
- 5.11 Current Phasors

Chapter 6. Capacitance in A.C. Circuits

- 6.1 Current Flow in a Capacitive Circuit

- 6.2 Capacitive Current Waveform
- 6.3 The Ohm's Law' for Capacitance
- 6.4 Capacitances in Parallel and in Series
- 6.5 Power in a Capacitive Circuit
- 6.6 Phasor Diagram for Capacitive Circuit
- 6.7 Capacitance and Resistance in Series
- 6.8 Impedance
- 6.9 Power in Mixed Circuits
- 6.10 Capacitance and Resistance in Parallel

Chapter 7. Inductance in A.C. Circuits

- 7.1 Current Flow in an Inductive Circuit
- 7.2 The Ohm's Law' for Inductance
- 7.3 Inductances in Series and in Parallel
- 7.4 Power in an Inductive Circuit
- 7.5 Phasor Diagram for Inductive Circuit
- 7.6 Inductance and Resistance in Series
- 7.7 Inductance and Resistance in Parallel
- 7.8 Transformers
- 7.9 Load Currents
- 7.10 Transformer Losses
- 7.11 Impedance Transformation

Chapter 8. The Tuned Circuit

- 8.1 Inductance and Capacitance in Series
- 8.2 L, C and R all in Series
- 8.3 The Series Tuned Circuit
- 8.4 Magnification
- 8.5 Resonance Curves
- 8.6 Selectivity
- 8.7 Frequency of Resonance
- 8.8 L and C in Parallel
- 8.9 The Effect of Resistance
- 8.10 Dynamic Resistance
- 8.11 Parallel Resonance
- 8.12 Frequency of Parallel Resonance
- 8.13 Series and Parallel Resonance Compared
- 8.14 The Resistance of the Coil
- 8.15 Dielectric Losses
- 8.16 H.F. Resistance
- 8.17 Cavity Resonators

- Chapter 9. Diodes
  - 9.1 Electronic Devices
  - 9.2 Diodes
  - 9.3 Thermionic Emission of Electrons
  - 9.4 The Vacuum Diode Valve
  - 9.5 Semiconductors
  - 9.6 Holes
  - 9.7 Intrinsic Conduction
  - 9.8 Effects of Impurities
  - 9.9 P-N Junctions
  - 9.10 The Semiconductor Diode
  - 9.11 Diode Characteristics
  - 9.12 Recapitulation
- Chapter 10. Triodes
  - 10.1 The Vacuum Triode Valve
  - 10.2 Amplification Factor
  - 10.3 Mutual Conductance
  - 10.4 Anode Resistance
  - 10.5 Alternating Voltage at the Grid
  - 10.6 Grid Bias
  - 10.7 The Transistor
  - 10.8 Transistor Characteristic Curves
  - 10.9 Transistor Parameters
  - 10.10 Field-Effect Transistors
  - 10.11 Insulated-Gate F.E.Ts
  - 10.12 Light-Sensitive Diodes and Triodes
- Chapter 11. The Triode at Work
  - 11.1 Input and Output
  - 11.2 Source and Load
  - 11.3 Feeds and Signals
  - 11.4 Load Lines
  - 11.5 Voltage Amplification
  - 11.6 An Equivalent Generator
  - 11.7 Calculating Amplification
  - 11.8 The Maximum-Power Law
  - 11.9 Transistor Load Lines
  - 11.10 Class A, B and C Operation
- Chapter 12. Transistor Equivalent Circuits
  - 12.1 The Equivalent Current Generator
  - 12.2 Duality

  - 12.3 Voltage or Current Generator?
  - 12.4 Transistor Equivalent Output Circuit
  - 12.5 Some Box Tricks
  - 12.6 Input Resistance
  - 12.7 Complete Transistor Equivalent Circuit
  - 12.8 A Simpler Equivalent
  - 12.9 Other Circuit Configurations: Common Collector
  - 12.10 Common Base
  - 12.11 A Summary of Results
  - 12.12 F.E.Ts and Valves
- Chapter 13. The Working Point
  - 13.1 Feed Requirements
  - 13.2 Effect of Amplification Factor Variations
  - 13.3 Influence of Leakage Current
  - 13.4 Methods of Base Biasing
  - 13.5 Biasing the F.E.T.
  - 13.6 Valve Biasing
  - 13.7 Biasing by Diodes
- Chapter 14. Oscillation
  - 14.1 Generators
  - 14.2 The Oscillatory Circuit
  - 14.3 Frequency of Oscillation
  - 14.4 Damping
  - 14.5 Maintaining Oscillation
  - 14.6 Practical Oscillator Circuits
  - 14.7 Resistance-Capacitance Oscillators
  - 14.8 Negative Resistance
  - 14.9 Amplitude of Oscillation
  - 14.10 Automatic Biasing of Oscillators
  - 14.11 Distortion of Oscillation
  - 14.12 Constancy of Frequency
  - 14.13 Crystal Oscillators
  - 14.14 Microwave Oscillators
- Chapter 15. Modulation
  - 15.1 The Need for Modulation
  - 15.2 Amplitude Modulation
  - 15.3 Methods of Amplitude Modulation
  - 15.4 Frequency Modulation
  - 15.5 Telegraphy and Keying
  - 15.6 Methods of Frequency Modulation

15.7 Sources of Modulating Signal
15.8 Theory of Sidebands
15.9 Channel Separation
15.10 Multiplex
15.11 Pulse Modulation Systems
Chapter 16. Transmission Lines
16.1 Feeders
16.2 Circuit Equivalent of a Line
16.3 Characteristic Resistance
16.4 Waves along a Line
16.5 Wave Reflection
16.6 Standing Waves
16.7 Line Impedance Variations
16.8 The Quarter-Wave Transformer
16.9 Fully Resonant Lines
16.10 Lines as Tuned Circuits
16.11 Waveguides or Radio-wave Plumbing
Chapter 17. Radiation and Antennas
17.1 Bridging Space
17.2 The Quarter-Wave Resonator Again
17.3 A Rope Trick
17.4 Electromagnetic Waves
17.5 Radiation
17.6 Polarization
17.7 Antennas
17.8 Radiation Resistance
17.9 Directional Characteristics
17.10 Reflectors and Directors
17.11 Antenna Gain
17.12 Choice of Frequency
17.13 Influence of the Atmosphere
17.14 Earthed Antennas
17.15 Feeding the Antenna
17.16 Tuning
17.17 Effective Height
17.18 Microwave Antennas
17.19 Inductor Antennas
Chapter 18. Detection
18.1 The Need for Detection
18.2 The Detector

18.3 Rectifiers
18.4 Linearity of Rectification
18.5 Rectifier Resistance
18.6 Action of Reservoir Capacitor
18.7 Choice of Component Values
18.8 The Diode Detector in Action
18.9 The Detector as a Load
18.10 Filters
18.11 Detector Distortion
18.12 Shunt-Diode Detector
18.13 Television Diode Detector
18.14 F.M. Detectors
18.15 Foster-Seeley Discriminator
18.16 Ratio Detector
18.17 Quadrature Detector
18.18 Phase-Locked-Loop Detector
18.19 Counter Discriminator
Chapter 19. Audio-Frequency Amplification
19.1 Recapitulation
19.2 Decibels
19.3 Gain/Frequency Distortion
19.4 Non-Linearity Distortion
19.5 Generation of Harmonics
19.6 Intermodulation
19.7 Allowable Limits of Non-Linearity
19.8 Phase Distortion
19.9 Loudspeakers
19.10 The Output Stage
19.11 Class B Amplification
19.12 Distortion with Class B
19.13 Class B Circuits
19.14 Compound Transistors
19.15 Negative Feedback
19.16 Phase Shift with Feedback
19.17 Input and Output Resistance
19.18 Linearising the Input
19.19 Some Circuit Details
19.20 Noise
19.21 Another Box Trick

Chapter 20. Selectivity and Tuning
- 20.1 Need for R.F. Amplification
- 20.2 Selectivity and Q
- 20.3 Over-Sharp Tuning
- 20.4 A General Resonance Curve
- 20.5 More Than One Tuned Circuit
- 20.6 Coupled Tuned Circuits
- 20.7 Effects of Varying Coupling
- 20.8 Practical Tuning Difficulties

Chapter 21. The Superheterodyne Principle
- 21.1 A Difficult Problem Solved
- 21.2 The Frequency-Changer
- 21.3 Frequency-Changers as Modulators
- 21.4 Types of Frequency-Changer
- 21.5 Conversion Conductance
- 21.6 Ganging the Oscillator
- 21.7 Whistles

Chapter 22. Radio-Frequency and Intermediate-Frequency Amplification
- 22.1 Amplification So Far
- 22.2 Active-Device Interelectrode Capacitances
- 22.3 Miller Effect
- 22.4 High-Frequency Effects in Transistors
- 22.5 Transistor Phasor Diagrams
- 22.6 Limiting Frequencies
- 22.7 An I.F. Amplifier
- 22.8 R.F. Amplification
- 22.9 Screening
- 22.10 Cross-Modulation
- 22.11 Automatic Gain Control
- 22.12 The Antenna Coupling
- 22.13 Klystrons and Travelling-Wave Tubes

Chapter 23. Cathode-Ray Tubes: Television and Radar
- 23.1 Description of Cathode-Ray Tube
- 23.2 Electric Focusing and Deflection
- 23.3 Magnetic Deflection and Focusing
- 23.4 Oscilloscopes
- 23.5 Time Bases
- 23.6 Application to Television
- 23.7 Characteristics of Television Signals
- 23.8 Television Receivers

- 23.9 Synchronizing Circuits
- 23.10 Scanning Circuits
- 23.11 Colour Television
- 23.12 Application to Radar

Chapter 24. Non-Sinusoidal Signal Amplification

- 24.1 Waveforms
- 24.2 Frequency and Time
- 24.3 Frequency Range
- 24.4 Importance of Phase Response
- 24.5 Obtaining the High-Frequency Response
- 24.6 Operational Amplifiers
- 24.7 Integrated Circuits

Chapter 25. Electronic Waveform Generators and Switches

- 25.1 Differentiating and Integrating Circuits
- 25.2 Waveform Shapers
- 25.3 Waveform Generators: Relaxation Oscillators
- 25.4 The Blocking Oscillator
- 25.5 The Multivibrator
- 25.6 Bistables
- 25.7 Monostables
- 25.8 Clampers and Clippers
- 25.9 Gates and Choppers
- 25.10 Design

Chapter 26. Computers

- 26.1 Analogue and Digital
- 26.2 Analogue Computers
- 26.3 Digital Computers
- 26.4 The Binary Scale
- 26.5 Logic Circuits
- 26.6 Counters
- 26.7 Stores
- 26.8 Computer Languages
- 26.9 Input and Output Devices
- 26.10 Microprocessors and Microcomputers

Chapter 27. Power Suppliers

- 27.1 The Power Required
- 27.2 Batteries
- 27.3 D.C. from A.C.
- 27.4 Types of Rectifier
- 27.5 Rectifier Circuits

- 27.6 Filters
- 27.7 Decoupling
- 27.8 Bias Supplies
- 27.9 Stabilisation
- 27.10 E.H.T.
- 27.11 Switch-Mode Power Suppliers
- 27.12 Cathode Heating

Appendix A: Algebraic Symbols

- A.1 Letter Symbols
- A.2 What Letter Symbols Really Mean
- A.3 Some Other Uses of Symbols
- A.4 Abbreviations
- A.5 How Numbers are Used

Appendix B: Graphs

- B.1 What is a Graph?
- B.2 Scales
- B.3 What a 'Curve' Signifies
- B.4 Three-Dimensional Graphs
- B.5 Significance of Slope
- B.6 Non-Uniform Scales

Appendix C: Alternative Technical Terms

Appendix D: Symbols and Abbreviations

Appendix E: Decibel Table

Index (1/3)

Index (2/3)

Index (3/3)

ELBS

Foundations of Wireless and Electronics

M. G. Scroggie

10th Edition

LOW-PRICED EDITION

Foundations of Wireless and Electronics

Tenth Edition

M. G. Scroggie
B.Sc., C.Eng., F.I.E.E.

Assisted in this revision by

S. W. Amos
B.Sc., C.Eng., M.I.E.E.

ELBS

English Language Book Society
and
Newnes Technical Books

Newnes Technical Books
is an imprint of the Butterworth Group
which has principal offices in
London, Boston, Durban, Singapore, Sydney, Toronto, Wellington

Originally published as 'Foundations of Wireless', 1936
by Iliffe Books
Eighth edition, rewritten and published as
'Foundations of Wireless and Electronics', 1971
Reprinted 1973
Ninth edition 1975 by Newnes Technical Books
Reprinted 1976, 1978, 1980
Tenth edition 1984

ELBS edition first published 1965
ELBS edition of ninth edition 1976
Reprinted 1978
ELBS edition of tenth edition 1984

© M. G. Scroggie, 1984

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, including photocopying and recording without the written permission of the copyright holder, application for which should be addressed to the Publishers. Such written permission must also be obtained before any part of this publication is stored in a retrieval system of any nature.

This book is sold subject to the Standard Conditions of Sale of Net Books and may not be resold in the UK below the net price given by the Publishers in their current price list.

British Library Cataloguing in Publication Data

Scroggie, M. G.
Foundations of wireless and electronics. — 10th ed.
1. Radio — Apparatus and supplies
I. Title II. Amos, S. W.
621.3841 TK6560

ISBN 0 408 01541 1

Typeset by Tunbridge Wells Typesetting Services Ltd
Printed in England by Cox & Wyman Ltd., Reading

Contents

Preface to Ninth Edition

Preface to Tenth Edition

1 General View of a System 1

1.1 What Wireless Does 1
1.2 Nature of Sound Waves 1
1.3 Characteristics of Sound Waves 3
1.4 Frequency 5
1.5 Wavelength 6
1.6 The Sender 7
1.7 The Receiver 7
1.8 Electrical Communication by Wire 8
1.9 Electric Waves 11
1.10 Why High Frequencies are Necessary 12
1.11 Radio Telegraphy 13
1.12 Tuning 14
1.13 Radio Telephony 15
1.14 Recapitulation 16

2 Electricity and Circuits 18

2.1 Electrons 18
2.2 Electric Charges and Currents 19
2.3 Conductors and Insulators 20
2.4 Electromotive Force 21
2.5 Electrical Units 22
2.6 Ohm's Law 23
2.7 Larger and Smaller Units 25
2.8 Circuit Diagrams 25
2.9 Resistances in Series and in Parallel 30
2.10 Series-Parallel Combinations 31
2.11 Resistance Analysed 33
2.12 Conductance 35
2.13 Kirchhoff's Laws 35
2.14 P.D. and E.M.F. 36
2.15 Electrical Effects 39
2.16 Instruments for Measuring Electricity 39
2.17 Electrical Power 41
2.18 A Broader View of Resistance 43

16 Transmission Lines 259

16.1 Feeders 259
16.2 Circuit Equivalent of a Line 259
16.3 Characteristic Resistance 262
16.4 Waves along a Line 264
16.5 Wave Reflection 266
16.6 Standing Waves 267
16.7 Line Impedance Variations 268
16.8 The Quarter-Wave Transformer 269
16.9 Fully Resonant Lines 270
16.10 Lines as Tuned Circuits 271
16.11 Waveguides or Radio-wave Plumbing 272

17 Radiation and Antennas 274

17.1 Bridging Space 274
17.2 The Quarter-Wave Resonator Again 274
17.3 A Rope Trick 276
17.4 Electromagnetic Waves 278
17.5 Radiation 279
17.6 Polarization 280
17.7 Antennas 281
17.8 Radiation Resistance 282
17.9 Directional Characteristics 282
17.10 Reflectors and Directors 284
17.11 Antenna Gain 286
17.12 Choice of Frequency 287
17.13 Influence of the Atmosphere 288
17.14 Earthed Antennas 290
17.15 Feeding the Antenna 291
17.16 Tuning 293
17.17 Effective Height 294
17.18 Microwave Antennas 294
17.19 Inductor Antennas 295

18 Detection 297

18.1 The Need for Detection 297
18.2 The Detector 297
18.3 Rectifiers 298
18.4 Linearity of Rectification 300
18.5 Rectifier Resistance 301
18.6 Action of Reservoir Capacitor 302
18.7 Choice of Component Values 304
18.8 The Diode Detector in Action 305
18.9 The Detector as a Load 306

Preface to Ninth Edition

Wireless—or radio—has branched out and developed so tremendously that very many books would be needed to describe it all in detail. In ordinary conversation, 'radio' has come to mean sound broadcasting, as distinct from 'TV' which gives pictures as well as sound, but of course the pictures come by radio just as much as the sound. Besides broadcasting sound and vision, radio is used for communication with and between ships, cars, aircraft, satellites and spacecraft; for direction-finding and radar (radiolocation), photograph and 'facsimile' transmission, telegraph and telephone links, meteorological probing of the upper atmosphere, astronomy, and other things. Very similar techniques are applied on an increasing scale to industrial control ('automation') and scientific and financial computation. All of these are based on the same foundational principles now generally comprised in the word *electronics.* The purpose of this book is to start at the beginning and lay these foundations, on which more detailed knowledge can then be built.

At the beginning . . . . If you had to tell somebody about a cricket match you had seen, your description would depend very much on whether or not your hearer was familiar with the jargon of the game. If he wasn't and you assumed he was, he would be puzzled. If it was the other way about, he would be irritated. There is the same dilemma with electronics. It takes much less time to explain it if the reader is familiar with methods of expression, such as symbols, that are taken for granted in technical discussion but not in ordinary conversation. This book assumes hardly any special knowledge. But if the use of graphs and symbols had to be completely excluded, or else accompanied everywhere by digressions explaining them, it would be very boring for the initiated. So the methods of technical expression are explained separately, as indicated below.

Then there are the technical terms. They are explained one by one as they occur and their first occurrence is distinguished by printing in italics; but in case any are forgotten they can be looked up in the Index; and the symbols and abbreviations in Appendix D. These references are there to be used whenever the meaning of anything is not understood.

Most readers find purely abstract principles very dull; it is more stimulating to have in mind some application of those principles. As

Preface to Tenth Edition

Since the ninth edition was published in 1975 there have been a number of significant developments in electronics. For example small thermionic valves are no longer used in new electronic equipment and manufacture of replacement types has ceased. The miniaturisation of integrated circuits has continued and a single i.c. can now contain more than 100 000 components, making even complex equipment very compact, as instanced by digital watches and domestic computers no larger than a book. Nearly all homes in this country now have colour television sets, but sales of plain (i.e. black-and-white) receivers continue mainly for use as second or portable sets.

To bring the book up to date most references to valves have been removed but sufficient has been left to provide an introduction to cathode-ray tubes and microwave tubes. Some information on senders has been omitted but more space has been devoted to modulation, including modern pulse methods. Accounts of up-to-date f.m. detectors have been added. Information on colour television has been expanded and now covers all the basic processes of transmission and reception. A new chapter on non-sinusoidal signal amplification has been introduced, and the chapter on computers has been updated.

All the circuit diagrams have been redrawn using the graphical symbols recommended by B.S.I.

This new edition was prepared by S.W. Amos who collaborated with M.G. Scroggie to ensure continuity of approach and style.

1983 S. W. AMOS

CHAPTER 1

General View of a System

Most of the chapters of this book are about subjects the usefulness of which, taken by themselves, might not be at all obvious. Proverbially, one might not be able to see the wood for the trees. So before examining these things in detail we may find it helpful to take a general view of a complete electronic system. The choice is a broadcasting and receiving system, because it is so familiar in everyday life and it happens to embody most of the principles of electronics.

1.1 What Wireless Does

People who remark on the wonders of wireless seldom seem to consider the fact that all normal persons can broadcast speech and song merely by using their voices. We can instantly communicate our thoughts to others, without wires or any other visible lines of communication and without even any sending or receiving apparatus outside of ourselves. If anything is wonderful, that is. Wireless, or radio, is merely a device for increasing the range.

There is a fable about a dispute among the birds as to which could fly the highest. The claims of the wren were derided until, when the great competition took place and the eagle was proudly outflying the rest, the wren took off from his back and established a new altitude record.

It has been known for centuries that communication over vast distances of space is *possible*—it happens every time we look at the stars and detect light coming from them. We know, then, that a long-range medium of communication exists. Radio simply uses this medium for the carrying of sound and vision.

Simply?

Well, there actually is quite a lot to learn about it. But the same basic principles can be applied to many other electronic systems.

1.2 Nature of Sound Waves

Radio (if, for convenience, we adopt the illogical convention that it excludes television), is as we have just observed, essentially a means of extending natural communication beyond its limits of range. It plays the part of the swift eagle in carrying the wren of human speech to remote distances. So first of all let us see what is required for 'natural' communication.

boys—representing the medium of communication—have all been standing still except for a small temporary displacement.

In a similar manner the cricket ball gives the bat a sudden push; the bat pushes the air next to it; that air gives the air around it a push; and so a wave of compression travels outwards in all directions like an expanding bubble. Yet all that any particular bit of air does is to move a small fraction of an inch away from the cause of the disturbance and back again. As the original impulse is spread over an ever-expanding wave-front the extent of this tiny air movement becomes less and less and the sound fainter and fainter.

1.3 Characteristics of Sound Waves

The generation of sound is only incidental to the functioning of cricket bats, but the human vocal organs make up a generator capable of emitting a great variety of sounds at will. What surprises most people, when the nature of sound waves is explained to them, is how such a variety of sound can be communicated by anything so simple as to-and-fro tremors in the air. Is it really possible that the subtle inflexions of the voice by which an individual can be identified, and all the endless variety of music, can be represented by nothing more than that?

Examining the matter more closely we find that this infinite variety of sounds is due to four basic characteristics, which can all be illustrated on a piano. They are:

(a) *Amplitude.* A single key can be struck either gently or hard, giving a soft or loud sound. The harder it is struck, the more violently the piano string vibrates and the greater the tremor in the air, or, to use technical terms, the greater the amplitude of the sound wave. Difference in amplitude is shown graphically in Fig. 1.2*a*. (Note that to-and-fro displacement of the air is represented here by up-and-down displacement of the curve, because horizontal displacement is usually used for time.)

(b) *Frequency.* Now strike a key nearer the right-hand end of the piano. The resulting sound is distinguishable from the previous one by its higher pitch. The piano string is tighter and lighter, so vibrates more rapidly and generates a greater number of sound waves per second. In other words, the waves have a higher *frequency,* which is represented in Fig. 1.2*b*.

(c) *Waveform.* If now several keys are struck simultaneously to give a chord, they blend into one sound which is richer than any of the single notes. The first three waves in Fig. 1.2*c* represent three such notes singly; when the displacements due to these separate notes are added together they give the more complicated waveform shown on the last line. Actually any one piano note itself has a complex waveform, which enables it to be distinguished from notes of the

Mathematically, Envelope is covered by Waveform, so may not always be listed separately.

The details of human and other sound generators are outside the scope of this book, but the sound waves themselves are the things that have to be carried by long-range communication systems such as radio. Even if something other than sound is to be carried (pictures, for instance) the study of sound waves is not wasted, because the basic characteristics illustrated in Fig. 1.2 apply to all waves—including radio waves.

The first (amplitude) is simple and easy to understand. The third and fourth (waveform and envelope) lead us into complications that are too involved for this preliminary survey, and will have to be gone into later. The second (frequency) is very important indeed and not too involved to take at this stage.

1.4 Frequency

Let us consider the matter as exemplified by the vibrating string or wire shown in Fig. 1.3. The upper and lower limits of its vibration are indicated (rather exaggeratedly, perhaps) by the dashed lines. If a point on the string, such as A, were arranged to record its movements on a strip of paper moving rapidly from left to right, it would trace out a wavy line which would be its displacement/time graph, like those in Fig. 1.2. Each complete up-and-down-and-back-again movement is called a *cycle*. In Fig. 1.4, where several cycles are shown, one has been picked out in heavy line. The time it takes, marked *T*, is known as its *period*.

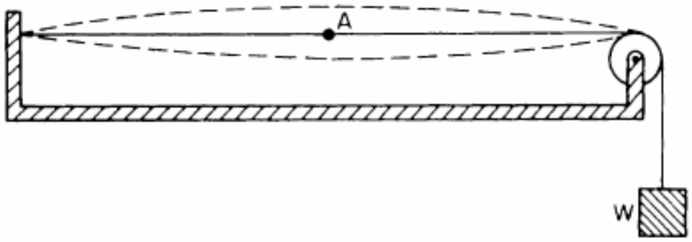


Fig. 1.3 A simple sound generator, consisting of a string held taut by a weight. When plucked it vibrates as indicated by the dashed lines

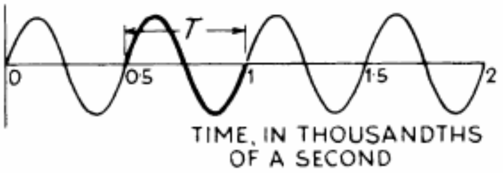


Fig. 1.4 The extent of one cycle in a series of waves is shown in heavy line. The time occupied by its generation, or by its passing a given point, is marked *T*

We have already noted that the rate at which the vibrations take place is called the *frequency*, and determines the pitch of the sound produced. Its symbol is f, and it used to be reckoned in cycles per second (abbreviated c/s) but the cycle per second is now called the hertz (abbr. Hz) in honour of Heinrich Hertz for a reason to be given very soon. Middle C in music has a frequency of 261 Hz. The full range* of a piano is 27 to 3516 Hz. The frequencies of audible sounds—about 20—20 000 Hz for young people with normal hearing but much lower for old people—are distinguished as *audio frequencies.*

You may have noticed in passing that T and f are closely related. If the period of each cycle is one-hundredth of a second, then the frequency is obviously one hundred cycles per second. Putting it in general terms, $f = 1/T$. So if the time scale of a waveform graph is given, it is quite easy to work out the frequency. In Fig. 1.4 the time scale shows that T is 0.005 s, so the frequency must be 2000 Hz.

1.5 Wavelength

Next, consider the air waves set up by the vibrating string. Fig. 1.5 shows an end view, with point A vibrating up and down. It pushes the air alternately up and down, and these displacements travel outwards from A. Places where the air has been temporarily moved a little farther from A than usual are indicated by full lines, and places where it is nearer by dashed lines; these lines should be imagined as expanding outwards at a speed of about 335 metres per second. So if, for example, the string is vibrating at a frequency of 20 Hz, at one

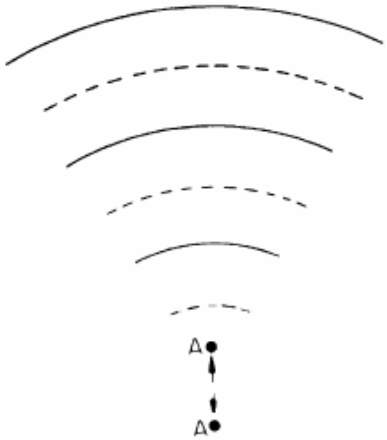


Fig. 1.5 End view of the vibrating string at A in Fig. 1.3. As it moves up and down over the distance AA it sends out air waves which carry some of its energy of vibration to the listener's ear

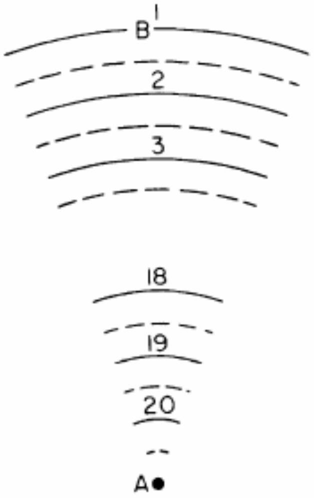


Fig. 1.6 Twenty successive waves from the string shown in Figs. 1.3 and 1.5

---

* In radio the word 'range' means distance that can be covered, so to avoid confusion a range of frequencies is more often referred to as a *band*.

second from the start the first air wave will be 335 metres away, and there will be 20 complete waves spread over that distance (Fig. 1.6). The length of each wave is therefore one-twentieth of 335 metres, which is 16.75 metres.

Now it is a fact that sound waves of practically all frequencies and normal amplitudes travel through the air at the same speed, so the higher the frequency the shorter the wavelength. This relationship can be expressed by the equation

$$\lambda = \frac{v}{f}$$

where the Greek letter λ (pronounced 'lambda') stands for wavelength, v for the velocity of the waves, and f for the frequency. If v is given in metres per second and f in hertz, then λ will be in metres. The letter v has been put here instead of 335, because the exact velocity of sound waves in air depends on its temperature; moreover the equation applies to waves in water, wood, rock, or any other substance, if the v appropriate to the substance is filled in.

It should be noted that it is the frequency of the wave which determines the pitch, and that the wavelength is a secondary matter depending on the speed of the wave. That this is so can be shown by sending a sound of the same frequency through water, in which the velocity is 1432 metres per second; the wavelength is therefore more than four times as great as in air, but the pitch, as judged by the ear, remains the same as that of the shorter air wave.

1.6 The Sender

The device shown in Fig. 1.3 is a very simple sound transmitter or sender. Its function is to generate sound waves by vibrating and stirring up the surrounding air. In some instruments, such as violins, the stirring-up part of the business is made more effective by attaching the vibrating parts to surfaces which increase the amount of air disturbed.

The pitch of the note can be controlled in two ways. One is to vary the weight of the string, which is conveniently done by varying the length that is free to vibrate, as a violinist does with his fingers. The other is to vary its tightness, as a violinist does when tuning, or can be done in Fig. 1.3 by altering the weight W. To lower the pitch, the string is made heavier or slacker. As this is not a book on sound there is no need to go into this further, but there will be a lot to say (especially in Chapter 8) on the electrical equivalents of these two things, which control the frequency of radio waves.

1.7 The Receiver

What happens at the receiving end? Reviving our memory of the stone in the pond, we recall that the ripples made corks and other small flotsam bob up and down. Similarly, air waves when they strike an object try to make it vibrate with the same characteristics as their own. That is how we can hear sounds through a door or partition; the door is made to vibrate, and its vibrations set up a new lot of air waves on our side of it. When they reach our ears they strike the ear-drums; these vibrate and stimulate a very remarkable piece of receiving mechanism, which sorts out the sounds and conveys its findings via a multiple nerve to the brain.

1.8 Electrical Communication by Wire

To extend the very limited range of sound-wave communication it was necessary to find a carrier. Nothing served this purpose very well until the discovery of electricity. At first electricity could only be controlled rather crudely, by switching the current on and off; so spoken messages had to be translated into a code of signals before they could be sent, and then translated back into words at the receiving end. A simple electric telegraph consisted of a battery and a switch or 'key' at the sending end, some device for detecting the current at the receiving end, and a wire between to carry the current.

Most of the current detectors made use of the discovery that when some of the electric wire was coiled round a piece of iron the iron became a magnet so long as the current flowed, and would attract other pieces of iron placed near it. If the neighbouring iron was in the form of a flexible diaphragm held close to the ear, its movements towards and away from the iron-cored coil when current was switched on and off caused audible clicks even when the current was very weak.

To transmit speech and music, however, it was necessary to make the diaphragm vibrate in the same complicated way as the sound waves at the sending end. Since the amount of attraction varied with the strength of current, this could be done if the current could be made to vary in the same way as the distant sound waves. In the end this problem was solved by quite a simple device—the carbon microphone, which, with only details improved, is still the type used in telephones.

In its simplest form, then, a telephone consists of the equipment shown in Fig. 1.7. Although this invention extended the range of speech from metres to kilometres, every kilometre of line weakened the electric currents and set a limit to clear communication. So it was not until the invention of the electronic valve, making it possible to amplify the complicated current variations, that telephoning could be done over hundreds and even thousands of kilometres.

So far we have concentrated on the transmission of sound. Now let us consider how moving images can be sent instantaneously over distances and we shall see that the problem has something in screen being controlled by the video signal.

The foregoing information refers to signals for plain television (i.e., without colour). The main modifications needed for colour come in Sec. 23.11. By the way 'black and white' is a clumsy and inaccurate description for pictures without colour, and 'monochrome' is wholly wrong because it means 'one colour', whereas white is composed of every colour! So, recalling the old advertising slogan, 'penny plain, twopence coloured', why not 'plain'?

What has been said about plain television applies equally to colour television because in all colour television systems the reproduced image is basically a plain one with areas filled in by colour. In fact all the fine detail in a colour television picture is reproduced in black and white. The colour information is transmitted by additional signals incorporated in the video signal. More information about colour television is given in Sec. 23.11.

To summarise the last few paragraphs, moving and coloured pictures can be reproduced at a distant point by sending to that point the video signal representing the original scene. The video signal thus corresponds to the audio signal of a sound system, but is more complex and, as we shall see later, contains much higher frequencies. Video signals can be sent to a distant fixed receiver by telephone lines just as audio signals can, but some other means of transmission is needed for sending video and audio signals to ships, aircraft, man-made satellites or the moon.

1.9 Electric Waves

At this stage we can profitably think again of the process of unaided voice communication. When you talk to another person you do not have to transfer the vibrations of your vocal organs directly to the ear-drum of your hearer by means of a rod or other 'line'. What you do is to stir up waves in the air, and these spread out in all directions, shaking anything they strike, including ear-drums.

'Is it possible', experimenters might have asked, 'to stir up *electric* waves by any means, and if so would they travel over greater distances than sound waves?' Gradually scientists supplied the answer. Clerk Maxwell showed mathematically that electric waves were theoretically possible, and indeed that in all probability light, which could easily be detected after travelling vast distances, was an example of such waves. But they were waves of unimaginably high frequency, about 5×10^{14} Hz. A few years later the German experimenter Hertz actually produced and detected electric waves of much lower (but still very high) frequency, and found that they shared with sound the useful ability to pass through things that are opaque to light.

Unlike sound, however, they are not carried by the air. They travel equally well (if anything, better) where there is no air or any other material substance present. So the broadcasting expression 'on the air' (or, worse still, 'air waves' for radio waves) is misleading. If the experiment of Fig. 1.1 is repeated with a source of electric waves instead of sound waves, extracting the air makes no appreciable difference. We know, of course, that light and heat waves travel to us from the sun across 149 668 620 kilometres of empty space. What does carry them is a debatable question. It was named ether (or aether, to avoid confusion with the anaesthetic liquid ether), but experiments which ought to have given results if there had been any such thing just didn't. So scientists deny its existence. But it is very hard for lesser minds to visualise how electric waves, whose behaviour corresponds in so many respects with waves through a material medium, can be propagated by nothing.

There is no doubt about their speed, however. Light waves travel through space at about 300 000 (more precisely 299 792) kilometres per second. Other sorts of electric waves, such as X-rays, ultra-violet and radio waves, travel at the same speed, and differ only in frequency. Their length is generally measured in metres; so filling in the appropriate value of v in the formula in Sec. 1.5 we have, very nearly,

$$\lambda = \frac{300\,000\,000}{f}$$

If, for example, the frequency is 1 000 000 Hz, the wavelength is 300 metres.

1.10 Why High Frequencies are Necessary

Gradually it was discovered that electric waves of these frequencies, called radio waves, are capable of travelling almost any distance, even round the curvature of the earth to the antipodes.

Their enormous speed was another qualification for the duty of carrying messages; the longest journey in the world takes less than a tenth of a second. But attempts to stir up radio waves of the same frequencies as sound waves were not very successful. To see why this is so we may find it helpful to consider again how sound waves are stirred up.

If you try to radiate air waves by waving your hand you will fail, because the highest frequency you can manage is only a few hertz, and at that slow rate the air has time to rush round from the front to the back of your hand and vice versa instead of piling up and giving the surrounding air a push. If you could wave your hand at the speed of an insect's (or even a humming bird's) wing, then the air would have insufficient time to equalise the pressure and would be alternately compressed and rarefied at the front or back of your hand and so would generate sound waves. Alternatively, if your hand were as big as the side of a house, it could stir up air waves even if waved only a few times per second, because the air would have too far to go from front to back every time. The frequency of these air waves would be too low to be heard, it is true; but they could be detected by the rattling of windows and doors all over the neighbourhood.

The same principle applies to radio waves. But because of the vastly greater speed with which they can rush round, it is necessary for the aerial or antenna—which corresponds to the waving hand—to be correspondingly vaster. To radiate radio waves at the lower audible frequencies the antenna would have to be miles high, so one might just as well (and more conveniently) use it on the ground level as a telephone line.

Even if there had not been this difficulty, another would have arisen directly large numbers of senders had started to radiate waves in the audible frequency band. There would have been no way of picking out the one wanted; it would have been like a babel of giants.

That might have looked like the end of any prospects of radio telephony, but human ingenuity was not to be beaten. The solution arose by way of the simpler problem of radio telegraphy, so let us follow that way.

1.11 Radio Telegraphy

Hertz discovered how to stir up radio waves; we will not bother about exactly how, because his method is obsolete. Their frequency was what we would now call ultra-high; in the region of hundreds of millions of hertz, so their length was only a few centimetres. Such frequencies are used nowadays for radar. He also found a method of detecting them over distances of a few yards. With the more powerful senders and more sensitive receivers developed later, ranges rapidly increased, until in 1901 Marconi is believed to have signalled across the Atlantic. The various sorts of detectors that were invented from time to time worked by causing an electric current to flow in a local receiving circuit when radio waves impinged on the receiving antenna. Human senses are unable to respond to these 'wireless' waves directly, and even if they could, the frequency is usually too high for the ear and far too low for the eye—but the electric currents resulting from their detection can be used to produce audible or visible effects.

At this stage we have the radio counterpart of the simple telegraph. Its diagram, Fig. 1.10, more or less explains itself. The sending key turns on the sender, which generates a rapid succession of radio waves, radiated by an antenna. When these reach the receiver they operate the detector and cause an electric current to flow, which can make a noise in a telephone earpiece. By working the sending key according to the Morse code, messages can be transmitted. That is roughly the basis of radio telegraphy to this day, though some installations have been elaborated almost out of recognition, and print the messages on paper faster than the most expert typist.

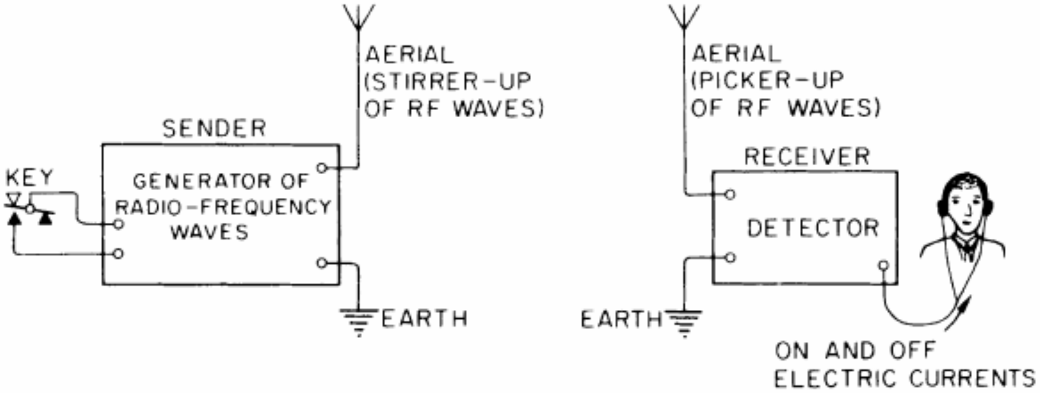


Fig. 1.10 Elementary radio telegraph. Currents flow in the receiving headphones and cause sound whenever the sender is radiating waves

1.12 Tuning

One important point to note is that the current coming from the detector does not depend on the frequency of the waves, within wide limits. (The frequency of the current is actually the frequency with which the waves are turned on and off at the sender.) So there is a wide choice of wave frequency. In Sec. 17.12 we shall consider how the choice is made; but in the meantime it will be sufficient to remember that if the frequency is low, approaching audibility, the antennas have to be immense; while if the fequency is much higher than those used by Hertz it is difficult to generate them powerfully, and they are easily obstructed, like light waves. The useful limits of these *radio* frequencies are about 15 000 to 50 000 000 000 Hz.

The importance of having many frequencies to choose from was soon apparent, when it was found to be possible—and highly advantageous—to make the receiver respond to the frequency of the sender and reject all others. This invention of *tuning* was essential to effective radio communication, and is the subject of a large part of this book. So it may be enough just now to point out that it has its analogy in sound. We can adjust the frequency of a sound wave by means of the length and tightness of the sender. A piano contains nearly a hundred strings of various fixed lengths and tightnesses, which we can select by means of the keys. If there is a second piano in the room, with its sustaining pedal held down so that all the strings are free to vibrate, and we strike a loud note on the first piano, the string tuned to the same note in the second can be heard to vibrate. If we had a single adjustable string, we could tune it to respond in this way to any one note of the piano and ignore those of substantially different frequency. In a corresponding manner the electrical equivalents of length (or weight) and tightness are adjusted by the tuning controls of a radio receiver to select the desired sending station.

1.13 Radio Telephony

The second important point is that (as one would expect) the strength of the current from the detector increases with the strength of the waves transmitted. So, with Fig. 1.7 in mind, we need no further hint to help us to convert the radio telegraph of Fig. 1.10 into a radio telephone or a broadcasting system, by substituting for the Morse key a device for varying the strength of the waves, just as the microphone varied the strength of the line current in Fig. 1.7. Such devices are called *modulators,* and will be referred to in more detail in Chapter 15. The microphone is still needed, because the output of the radio sender is not controlled directly by the impinging sound waves but by the wavily varying currents from the microphone.

A practical system needs amplifiers in various places, firstly to amplify the feeble sound-controlled currents from the microphone until they are strong enough to modulate a powerful wave sender; then to amplify the feeble radio-frequency currents stirred up in a

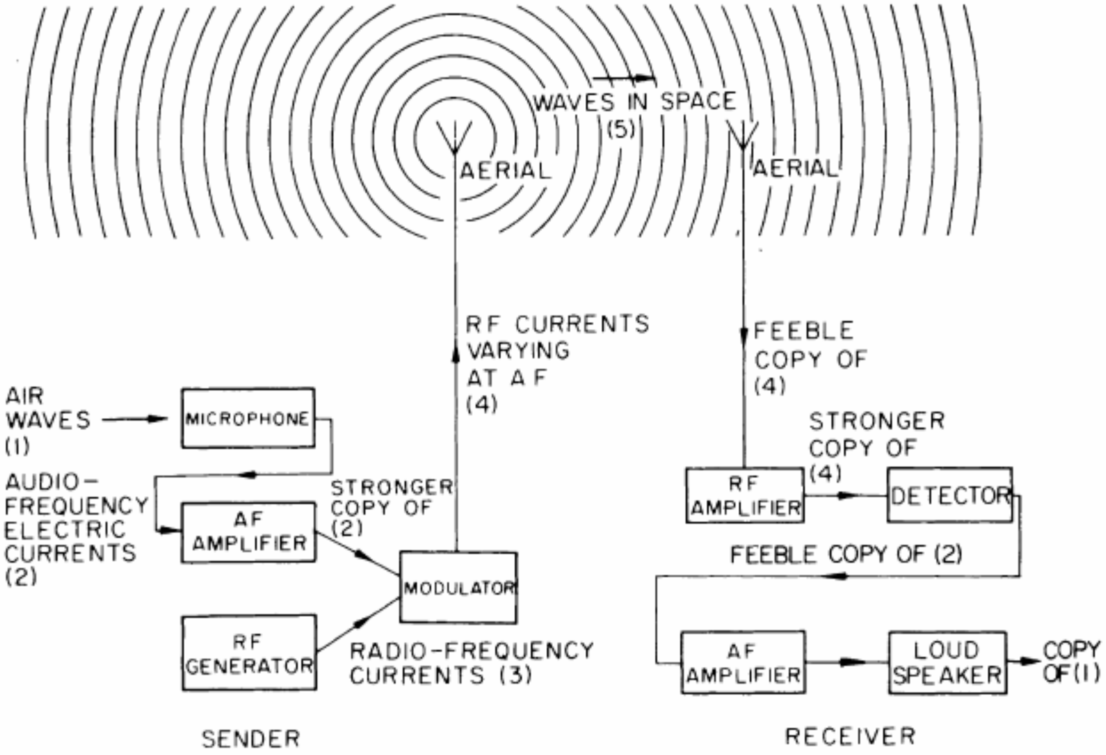

WAVES IN SPACE
(5)
AERIAL
AERIAL
RF CURRENTS VARYING AT A F (4)
FEEBLE COPY OF (4)
AIR WAVES (1)
MICROPHONE
STRONGER COPY OF (4)
AUDIO-FREQUENCY ELECTRIC CURRENTS (2)
STRONGER COPY OF (2)
AF AMPLIFIER
RF AMPLIFIER
DETECTOR
MODULATOR
FEEBLE COPY OF (2)
RF GENERATOR
RADIO-FREQUENCY CURRENTS (3)
AF AMPLIFIER
LOUD SPEAKER
COPY OF (1)
SENDER
RECEIVER

Fig. 1.11 The radio telegraph of Fig. 1.10 converted into a radio telephone by controlling the output of the sender by sound waves instead of just turning it on and off. In practice it is also generally necessary to use amplifiers where shown

distant receiving antenna by the far flung waves; then another to amplify the audio-frequency currents from the detector so that they are strong enough to work a loudspeaker instead of the less convenient headphones. This somewhat elaborated system is indicated—but still only in broadest outline—by the block diagram, Fig. 1.11.

We have already shown the analogy between audio and video signals and in order to obtain a simple block diagram of a television broadcasting system we can, in Fig. 1.11, replace the microphone by a television camera, the a.f. amplifier by a video amplifier and the loudspeaker by a picture tube. We must remember, however, that the camera tube and picture tube require sources of power for vertical and horizontal deflection of their electron beams and that a television system includes a sound system. So another system such as that illustrated in Fig. 1.11 is required to convey the sound associated with the picture.

1.14 Recapitulation

At this stage it may be as well to review the results of our survey. We realise that in all systems of communication the signals must, in the end, be detectable by the human senses. Of these, hearing and seeing are the only ones that matter. Both are wave-operated. Our ears respond to air waves between frequency limits of about 20 and at most 20 000 Hz; our eyes respond to waves within a narrow band of frequency centred on 5×10^{14} Hz—waves not carried by air but by space, and which have been proved to be electrical in nature.

Although our ears are remarkably sensitive, air-wave (i.e. sound) communication is very restricted in range. Eyes, it is true, can see over vast distances, but communication is cut off by the slightest wisp of cloud. If, however, we can devise means for 'translating' audible or visible waves into and out of other kinds of waves, we have a far wider choice, and can select those that travel best. It has been found that these are electric waves of lower frequencies than the visible ones, but higher than audible frequencies. The lowest of the radio frequencies require excessively large antennas to radiate them, and the highest are impeded by the atmosphere. All travel through space at the same speed of nearly 300 000 000 metres per second, so the wavelength in metres can be calculated by dividing that figure by the frequency in Hz.

The frequency of the sender is arranged by a suitable choice and adjustment of the electrical equivalents of weight and tightness (or, more strictly, slackness) in the tuning of sound generators. In the same way the receiver can be made to respond to one frequency and reject others. There are so many applicants for frequencies—broadcasting authorities, telegraph and telephone

CHAPTER 2

Electricity and Circuits

2.1 Electrons

The exact nature of electricity is a mystery that may never be fully cleared up, but from what is known it is possible to form a sort of working model or picture which helps us to understand how it produces the results it does, and even to think out how to produce new results. The reason why the very existence of electricity went unnoticed until comparatively recent history was that there are two opposite kinds which exist in equal quantities and, unless separated in some way, cancel one another out. This behaviour reminded the investigating scientists of the use of positive (+) and negative (–) signs in arithmetic: the introduction of either $+1$ or -1 has a definite effect, but $+1-1$ equals just nothing. So, although at that time hardly anything was known about the two kinds of electricity, they were called positive and negative. Both positive and negative electricity produce very remarkable effects when they are separate, but a combination of equal quantities shows no signs of electrification.

Further research led to the startling conclusion that all matter consists largely of electricity. All the thousands of different kinds of matter, whether solid, liquid, or gaseous, consist of atoms which contain only a very few different basic components. Of these we need take note of only one kind—particles of negative electricity, called *electrons.* For a simple study it is convenient to imagine each atom as a sort of ultra-microscopic solar system in which a number of electrons are distributed around a central nucleus, somewhat as our earth and the other planets around the sun (Fig. 2.1). (The subject is treated in much more detail in the author's *The Electron in Electronics*). The nucleus is generally a more or less composite structure; it makes up nearly all the weight of the atom, and the number and type of particles it contains determine which element it is—oxygen, carbon, iron, etc. It is possible to change some elements into others by breaking up the nucleus under very intense laboratory treatment. For our present purpose, however, the only things we need remember about the nucleus are that it is far heavier than the electrons, and that it has a surplus of positive electricity.

In the normal or unelectrified state of the atom this positive surplus is exactly neutralised by the planetary electrons. But it is a comparatively easy matter to dislodge one or more of these electrons from each atom. One method is by simply rubbing; for example, if a glass rod is rubbed vigorously with silk some of the electrons belonging to the glass are transferred to the silk. Doing this produces no change in the material itself—the glass is still glass and the silk remains silk. As separate articles they show no obvious evidence of the transfer. But if they are brought close together it is found that they attract one another. And the glass rod can pick up small scraps of paper. It is even possible, in a dry atmospherre, to produce sparks. These curious phenomena gradually fade away, and the materials become normal again.

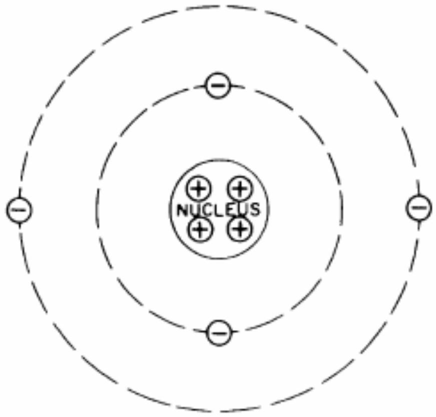


Fig. 2.1 Diagrammatic picture of an atom, showing the nucleus surrounded by a number of electrons

2.2 Electric Charges and Currents

The surplus of electrons on the silk is called a negative electric *charge*. The glass has a corresponding deficiency of electrons, so its positive nuclear electricity is incompletely neutralised, and the result is a positive charge. It is a pity that the people who decided which to call positive did not know anything about electrons, because if they had they would certainly have called electrons positive and so spared us a great deal of confusion. As it is, however, one just has to remember that a *surplus* of electrons is a *negative* charge.

Any unequal distribution of electrons is a condition of stress. Forcible treatment of some kind is needed in order to create a surplus or deficiency anywhere, and if the electrons get a chance they move back to restore the balance; i.e., from negatively to positively charged bodies. This tendency shows itself as an attraction between the bodies themselves. That is what is meant by saying that opposite charges attract.

The space between opposite charges, across which this attraction is exerted, is said to be subject to an *electric field*. The greater the charges, the more intense is the field and the greater the attractive force.

If a number of people are forcibly transferred from town A to town B, the pressure of overcrowding and the intensity of desire to leave B depend, not only on the number of displaced persons, but also on the size of B and possibly on whether there is a town C close by with plenty of accommodation. The electrical pressure also depends on other things than the charge, reckoned in displaced electrons; so it is more convenient to refer to it by another term, namely *difference of potential*, often abbreviated to p.d. We shall consider the exact relationship between charge and p.d. in the next chapter. The thing to remember now is that wherever there is a difference of potential between two points there is a tendency for electrons to move from the point of lower (or negative) potential to that of higher (or positive) potential in order to equalise the distribution.* If they are free to move they will do so; and their movement is what we call an *electric current.*

This is where we find it so unfortunate that the names 'positive' and 'negative' were allocated before anybody knew about electrons. For it amounted to a guess that the direction of current flow was the opposite to that in which we now know electrons flow. By the time the truth was known this bad guess had become so firmly established that reversing it would have caused worse confusion. Moreover, when the positively charged atoms (or positive *ions,* as they are called) resulting from the removal of electrons are free to do so they also move, in the direction + to –, though much more slowly owing to their greater mass. So this book follows the usual custom of talking about current flowing from + to – or high to low potential, but it must be remembered that most often this means electrons moving from — to + .

2.3 Conductors and Insulators

Electrons are not always free to move. Except in special circumstances (such as high temperature, which we shall consider in Chapter 9) atoms do not allow their electrons to fly off completely on their own, even when they are surplus. The atoms of some substances go so far as to allow frequent exchanges, however, like dancers in a Paul Jones, and in fact such exchanges go on all the time, even in the normal unelectrified state. The directions in which the electrons flit from one atom to another are then completely random, because there is nothing to influence them one way or another.

But suppose the whole substance is pervaded by an electric field.

---

*Note that + 1 is positive with respect to – 1 or even to 0, but it is negative with respect to + 2; – 1 or 0 are both positive with respect to – 2. There is nothing inconsistent in talking about two oppositely charged bodies one minute, and in the next referring to them both as positive (relative to something else).

The electrons feel an attraction towards the positive end; so between partnerships they tend to drift that way. The drift is what is known as an electric current, and substances that allow this sort of thing to go on are called *conductors* of electricity. All the metals are more or less good conductors; hence the extensive use of metal wire. Carbon and some liquids are fairly good conductors.

Note that although electrons start to go in at the negative end and electrons start to arrive at the positive end the moment the p.d. is set up, this does not mean that these are the same electrons, which have instantaneously travelled the whole length of the conductor. An electron drift starts almost instantaneously throughout miles of wire, but the speed at which they drift is seldom much more than an inch a minute.

Other substances keep their electrons, as it were, on an elastic leash which allows them a little freedom of movement, but never 'out of sight' of the atoms. If such a substance occupies the space between two places of different potential, the electrons strain at the leash in response to the positive attraction, but a continuous steady drift is impossible. Materials of this kind, called *insulators*, can be used to prevent charges from leaking away, or to form boundaries restricting currents to the desired routes. Dry air, glass, polythene, rubber, and paraffin wax are among the best insulating materials. None is absolutely perfect, however; electrified glass rods, for example, gradually lose their charges.

We shall see in Chapter 9 that there is a third class of substances called *semiconductors,* of immense importance in electronics.

2.4 Electromotive Force

If electrons invariably moved from – to +, as described, they would in time neutralise all the positive charges in the world, and that, for all practical purposes, would be the end of electricity. But fortunately there are certain appliances, such as batteries and dynamos (or generators) which can force electrons to go from + to –, contrary to their natural inclination. In this way they can continuously replenish a surplus of electrons. Suppose A and B in Fig. 2.2 are two insulated metal terminals, and a number of electrons have been transferred from A to B, so that A is at a higher potential than B. If now they are joined by a wire, electrons will drift along it, and, if that were all, the surplus would soon be used up and the potential difference would disappear. But A and B happen to be the terminals of a battery, and as soon as electrons leave B the battery provides more, while at the same time it withdraws electrons arriving at A. So while electrons are moving from B to A through the wire, the battery keeps them moving from A to B through itself. The battery is, of course, a conductor; but if it were only that it would be an additional path for electrons to go from B to A and dissipate the charge all the quicker. It is remarkable, then, for being able to make electrons move against a p.d. This ability is called *electromotive force* (usually abbreviated to e.m.f.).

Fig. 2.2 shows what is called a *closed circuit*, there being a continuous endless path for the current. This is invariably necessary for a continuous current, because if the circuit were opened anywhere, say by disconnecting the wire from one of the terminals, continuation of the current would cause electrons to pile up at the break, and the potential would increase without limit; which is impossible. So the electron flow ceases throughout the conducting path the moment the path is broken at any point.

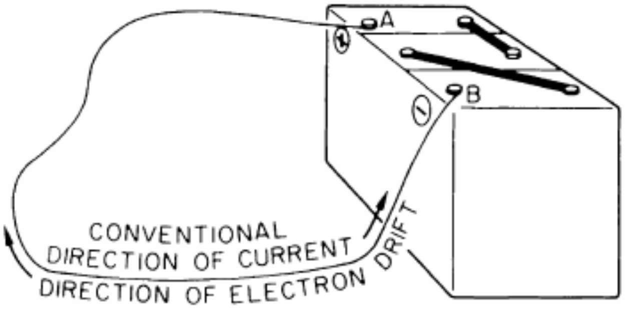

A
B
+
−
CONVENTIONAL DIRECTION OF CURRENT
DIRECTION OF ELECTRON DRIFT

Fig. 2.2 A very simple electrical circuit, round which the electromotive force of a battery makes a current of electrons flow continuously

2.5 Electrical Units

The amount or strength of an electric current might reasonably be reckoned as the number of electrons passing any point in the circuit each second, but the electron is so extremely small that such a unit would lead to inconveniently large numbers. For practical purposes it has been agreed to base the unit on the metric system. It is called the *ampere* (or more colloquially the amp), and happens to be about 6 240 000 000 000 000 000 (or 6.24×10^{18}) electrons per second.

As one would expect, the number of amps caused to flow in a given circuit depends on the strength of the electromotive force operating in it. The practical unit of e.m.f. is the *volt.*

Obviously the strength of current depends also on the circuit—whether it is made up of good or bad conductors. This fact is expressed by saying that circuits differ in their electrical *resistance.* The resistance of a circuit or of any part of it can be reckoned in terms of the e.m.f. required to drive a given current through it. For convenience the unit of resistance is made numerically equal to the number of volts required to cause one amp to flow, and to avoid the cumbersome expression 'volts per amp' this unit of resistance has been named the *ohm.*

By international agreement the following letter symbols have been allocated to denote these electrical quantities and their units, and will be used from now on:

Quantity	*Symbol for quantity*	*Unit of quantity*	*Symbol for unit*
E.M.F.	E	Volt	V
Current	I	Ampere	A
Resistance	R	Ohm	Ω

(Ω is a Greek capital letter omega)

These three important quantities are not all independent, for we have just defined resistance in terms of e.m.f. and current. Expressing it in symbols:

$$R = \frac{E}{I}$$

Note that sloping letters (italics) are used to denote quantities, and upright ones (roman) for the units. This is always done; see Sec. A.4.

2.6 Ohm's Law

The question immediately arises: does R depend only on the circuit, or does it depend also on the current or voltage? This was one of the first and most important investigations into electric currents. We can investigate it for ourselves if we have an instrument for measuring current, called an *ammeter,* a battery of identical cells, and a length of thin wire or other resistive conductor. A closed circuit is formed of the wire, the ammeter, and a varying number of the cells; and the ammeter reading corresponding to each number of cells is noted (Fig. 2.7). The results can be plotted in the form of a graph (Fig. 2.3).

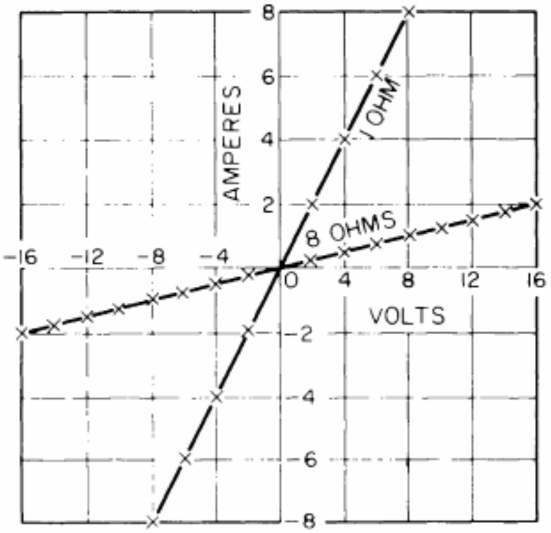


Fig. 2.3 Graph showing the results of an experiment on the relationship between current in amps and e.m.f. in volts, for two different circuits

had fallen to 900×0.007 = 6.3 V. Note that in driving current through a resistance an e.m.f. causes a p.d. of the same voltage between the ends of the resistance. This point is elaborated in Sec. 2.14.

2.7 Larger and Smaller Units

It is unusual to describe a current as 0.007 ampere, as was done just now; one speaks of '7 milliamperes', or, more familiarly, '7 milliamps'. A milliampere is one-thousandth part of an ampere. Several other prefixes are used; they are tabulated in Appendix D, the commonest being

Prefix	*Meaning*	*Symbol*
milli-	one thousandth of	m
micro-	one millionth of	μ
nano-	one thousandth-millionth of	n
pico	one billionth of (million millionth)	p
kilo-	one thousand	k
mega-	one million	M
giga	one thousand million	G

The prefixes can be put in front of any unit; one speaks commonly of milliamps, microamps, megohms, and so on. 'Half a megohm' is easier to say than 'Five hundred thousand ohms', and '$\frac{1}{2}$ MΩ' is quicker to write than '500 000 Ω'.

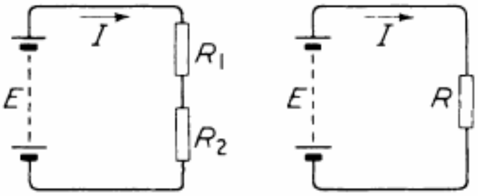
It must be remembered, however, that Ohm's law in the forms given on p.24 assumes volts, ohms, and amps. If a current of 5 mA is flowing through 15 000 Ω, the voltage across that resistance will *not* by 75 000 V. But since most of the currents we shall be concerned with are of the milliamp order it is worth noting that there is no need to convert them to amps if R is expressed in *thousands* of ohms (kΩ). So in the example just given one can get the correct answer, 75 V, by multiplying 5 by 15.

2.8 Circuit Diagrams

At this stage it will be as well to start getting used to circuit diagrams. In a book like this, concerned with general principles rather than with constructional details, what matters about (say) a battery is not its shape nor the design of its label, but its voltage and how it is connected in the circuit. So it is a waste of time drawing a picture as in Fig. 2.2. All one needs is the symbol shown as Fig. 2.4*a*, which represents one cell. An accumulator cell gives 2 V, and a dry cell

2.9 Resistances in Series and in Parallel

Complicated circuit networks can be tackled by successive stages of finding a single element that is equivalent (as regards the quantity to be found) to two or more. Consider first two resistances R_1 and R_2, in series with one another and with a source of e.m.f. E (Fig. 2.9*a*).



(a) (b)

Fig. 2.9 Resistance in series. The circuit (*b*) is equivalent to the circuit (*a*), in the sense that both take the same current from the battery, if $R = R_1 + R_2$

To bring this circuit as a whole within the scope of calculation by Ohm's law we need to find the single resistance R (Fig 2.9*b*) equivalent to R_1 and R_2 together.

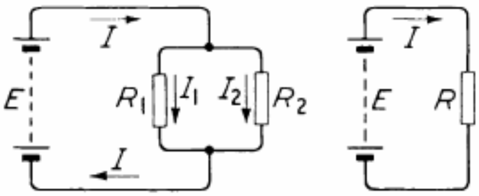
We know that the current in the circuit is everywhere the same; call it I. Then by Ohm's law, the voltage across R_1 is IR_1, and that across R_2 is IR_2. The total voltage across both must therefore be $IR_1 + IR_2$, or $I(R_1 + R_2)$, which must be equal to the voltage E.

In the equivalent circuit, E is equal to IR, and since to make the circuits truly equivalent, the currents must be the same in both for the same battery voltage, we see that

$$R = R_1 + R_2.$$

Generalising from this result, we conclude that: *The equivalent resistance of several resistances in series is equal to the sum of their individual resistances.* It follows that if a number of resistances are connected in series, the effective resistance is always greater than any individual resistance.

Turning to the parallel-connected resistances of Fig. 2.10*a*, we see that they have the same voltage across them; in this case the e.m.f. of the battery. Each of these resistances will take a current depending on its own resistance and on the e.m.f.— this simplest case of Ohm's



(a) (b)

Fig. 2.10 Resistance in parallel. The circuit (*b*) is equivalent to the circuit (*a*), in the sense that both take the same current from the battery, if $1/R = 1/R_1 + 1/R_2$

law. Calling the currents respectively I_1 and I_2, we therefore know that $I_1 = E/R_1$ and $I_2 = E/R_2$. The total current drawn is the sum of the two: it is

$$I = \frac{E}{R_1} + \frac{E}{R_2} = E\left(\frac{1}{R_1} + \frac{1}{R_2}\right)$$

In the equivalent circuit of Fig. 2.10*b* the current is E/R, which may also be written $E\ (1/R)$. Since, for equivalence between the circuits, the current must be the same for the same battery voltage, we see that

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2}$$

Generalising from this result, we conclude that: *The reciprocal of the equivalent resistance of several resistances in parallel is equal to the sum of the reciprocals of their individual resistances.* It follows from this that the effective resistance of a number of resistances in parallel is always less than the smallest of them.

If the resistances of Fig. 2.10*a* were 100 Ω and 200 Ω, the single resistance R which, connected in their place, would draw the same current is given by $1/R = 1/100 + 1/200 = 0.01 + 0.005 = 0.015$. Hence, $R = 1/0.015 = 66.67\ \Omega$. This could be checked by adding together the individual currents through 100 Ω and 200 Ω, and comparing the total with the current taken from the same voltage-source by 66.67 Ω. In both cases the result is 15 mA per volt of battery.

We can summarise the two rules in symbols:

1. Series Connection: $R = R_1 + R_2 + R_3 + R_4 + \ \ldots.$
2. Parallel Connection: $1/R = 1/R_I + 1/R_2 + 1/R_3 + \ \ldots$

For only two resistances in parallel, a more convenient form of the same rule can be obtained by multiplying above and below by R_1R_2:

$$R = \frac{1}{1/R_1 + 1/R_2} = \frac{R_1R_2}{R_1 + R_2} = \frac{\text{product of resistances}}{\text{sum of resistances}}$$

2.10 Series-Parallel Combinations

How the foregoing rules, derived directly from Ohm's law, can be applied to the calculation of more complex circuits can perhaps best be illustrated by a thorough working-out of one fairly elaborate network, Fig 2.11. We will find the total current flowing, the equivalent resistance of the whole circuit, and the voltage and current of every resistor individually.

The policy is to look for any resistances that are in simple series or parallel. The only two in this example are R_2 and R_3. Writing R_{23} to symbolise the combined resistance of R_2 and R_3 taken together, we show, across the whole complex system $R_2R_3R_4R_5$ in Fig. 2.11. Across R_5 there comes the whole of this voltage; the current through this resistor will therefore be 27.75/1000 A = 27.75 mA.

The same p.d. across R_{234} of Fig. 2.12*b*, or across the system $R_2R_3R_4$ of Fig 2.11, will drive a current of 27.75/292.8 = 94.75 mA through this branch. The whole of this flows through R_4 (2.12*a*), the voltage across which will accordingly be 150 × 0.09 475 = 14.21 V. Similarly, the p.d. across R_{23} in Fig. 2.12*a*, or across both R_2 and R_3 in Fig. 2.11, will be 0.09475 × 142.8 = 13.54 V, from which we find that the currents through R_2 and R_3 will be respectively 13.54/200 and 13.54/500 A, or 67.68 and 27.07 mA, making up the required total of 94.75 mA for this branch.

This completes an analysis of the entire circuit; we can now collect our scattered results in the form of the following table.

	Resistance (Ω)	*Current through it* (mA)	*Potential difference across it* (V)
R_1	100	122.5	12.25
R_2	200	67.68	13.54
R_3	500	27.07	13.54
R_4	150	94.75	14.21
R_5	1000	27.75	27.75
R	326.5	122.5	40

Note that by using suitable resistors any potential intermediate between those given by the terminals of the battery can be obtained. For instance, if the lower and upper ends of the battery in Fig. 2.11 are regarded as 0 and + 40 (they could equally be reckoned as – 40 and 0, or – 10 and + 30, with respect to any selected level of voltage), the potential of the junction between R_3 and R_4 is 14.21 V. The arrrangement is therefore called a *potential divider*, and is often employed for obtaining a desired potential not given directly by the terminals of the source. If a sliding connection is provided on a resistor, to give a continuously variable potential, it is generally known—though not always quite justifiably—as a *potentiometer.*

2.11 Resistance Analysed

So far we have assumed the possibility of almost any value of resistance without considering exactly what determines the resistance of any particular resistor or part of a circuit. We understand that different materials vary widely in the resistance they offer to the flow of electricity, and can guess that with any given material long piece will offer more resistance than a short piece, and a thin piece than a thick. We have indeed actually proved as should be as low as possible) and for the special alloys intended to have a high resistance.

Resistance varies to some extent with temperature, so the tables show the temperature at which they apply, and also the proportion by which the resistance increases with rise of temperature.

2.12 Conductance

It is often more convenient to work in terms of the *ease* with which a current can be made to flow, rather than the difficulty; in other words, in *conductance* rather than resistance. The symbol for conductance is G, and its unit (called the siemens, S; formerly, the mho) is equal to the number of amps caused to flow by one volt. So $I = EG$ is an alternative form of $I = E/R$; and $G = 1/R$.

It is not difficult to see that corresponding to the equation for calculating resistance there will be one for conductance:

$$G = \frac{\sigma A}{l}$$

where σ (Greek 'sigma') stands for *conductivity* (siemens per metre).

The advantage of conductance appears chiefly in parallel circuits, for which it is easy to see that the formula is $G = G_1 + G_2 + \ldots$

2.13 Kirchhoff's Laws

We have already taken it as obvious that no one point can be at more than one potential at the same time. (If it is not obvious, try to imagine a point with a difference of potential between itself; then electrons will flow from itself to itself, i.e., a journey of no distance and therefore non-existent.) So if you start at any point on a closed circuit and go right round it, adding up all the voltages on the way, with due regard for + and –, the total is bound to be zero. If it isn't, a mistake has been made somewhere; just as a surveyor must have made a mistake if he started from a certain spot, measuring his ascents as positive and his descents as negative, and his figures told him that when he returned to his starting-point he had made a net ascent or descent.

This simple principle is the modern form of one of what are known as Kirchhoff's laws, and is a great help in tackling systems of resistances where the methods already described fail because there are no two that can be simply combined. One writes down an equation for each closed loop or mesh, by adding together all the voltages and equating the total to zero, and one then solves the resulting simultaneous equations. That can be left for more advanced study, but the same law is a valuable check on any circuit calculations. Try applying it to the several closed loops in Fig. 2.11, such as that formed by E, R_1, and R_5. If we take the route clockwise we go 'uphill' through the battery, becoming more positive by 40 V. Coming down through R_1 we move from positive to negative so add -12.25 V, and, through R_5, -27.75 V, reaching the starting point again. Check: $40-12.25-27.75=0$.

Taking another clockwise route via R_3, R_5, and R_4, we get $13.54-27.75+14.21=0$. And so on for any closed loop.

The other Kirchhoff law is equally obvious; it states that if you count all currents arriving at any point in a circuit as positive, and those leaving as negative, the total is zero. This can be used as an additional check, or as the basis of an alternative method of solving circuit problems.

2.14 P.D. and E.M.F.

The fact that both e.m.f. and p.d. are measured in volts can be confusing. So it will be time well spent if, before going any farther, we look into the matter rather carefully.

We have just been likening difference of potential to difference in height. This analogy is implied whenever we talk about low and high potential. But although we measure height in feet or metres what we have in mind is not just a *distance*; we are really concerned with the force of gravity close to a massive body, usually the earth. ('Height' has no meaning in outer space.) This force tends to make things move downwards, or fall. Corresponding to this gravitational field is the electric field between opposite electric charges, which tends to make mobile charges move in the direction of that field, towards the low-potential or negative end if the mobile charges are positive ones, but 'upwards' if they are electrons. Just as an effort, which we might call a weight-motive force, is required to lift a weight against the force of a gravitational field, so an electromotive force is needed to lift positive electric charges to a point of higher potential; that is to say, against the force of an electric field. And just as the amount of weight-motive force could be reckoned as the height it could lift a weight—any weight—so e.m.f. is specified as the p.d. it can overcome. If water has to be supplied to the top floor of a 60m block of flats, then a water-motive force which could be specified as not less than 60 m must be provided by a pump or other active device. The difference in height that the pressure of the water source can overcome is the most directly informative way of stating it. Similarly the p.d. in volts that a source of electricity can overcome is the most helpful way of stating its e.m.f., which indeed is sometimes referred to as its pressure.

Another reason for reckoning e.m.f. in volts is that it cannot be measured as such, but only in terms of the p.d. it can overcome or create.

reserved for e.m.f., p.ds being denoted by V. But often there is no need to make any distinction, and we shall see in Sec. 5.11 that attempting to do so can lead to controversy and confusion, so there is much to be said for reckoning in voltage rises (falls being negative rises) as for example those indicated in Fig. 2.14*b*.

2.15 Electrical Effects

It is time now to see what electricity can *do*. One of the most familiar effects of an electric current is heat. It is as if the jostling of the electrons in the conductor caused a certain amount of friction. But whereas it makes the wire in an electric fire red hot, and in an electric light bulb white hot, it seems to have no appreciable effect on the flex and other parts of the circuit—in a reputable installation, anyway. Reasoning of the type we used in connection with Fig. 2.13 shows that the rate at which heat is generated in a conductor containing no e.m.fs is proportional to the product of the current flowing through it and the voltage across it (i.e., the p.d. between its ends); in symbols, EI. Since both E and I are related to the resistance of the conductor, we can substitute IR for E or E/R for I, getting I^2R or E^2/R as alternative measures of the heating effect. Thus for a given current the heating is proportional to the resistance. The wires connecting electric lamps and heaters to the generator are chosen to have a resistance low compared with that of the electric devices themselves. I^2R shows that the heating increases very rapidly as the current rises. Advantage is taken of this in fuses which are short pieces of wire that melt (or 'blow') and cut off the current if, owing to a short-circuit somewhere, it rises dangerously above normal.

Another effect of an electric current is to magnetise the space around it. This effect is particularly strong when the wire is coiled round a piece of iron. As we shall see, it is of far more significance then merely being a handy way of making magnets.

An effect that need not concern us much is the production of chemical changes, especially when the current is passed through watery liquids. A practical example is the charging of batteries. The use of the word 'charging' for this process is unfortunate, for no surplus or deficiency of electrons is accumulated by it.

All these three effects are reversible; that is to say, they can be turned into causes, the effect in each case being an e.m.f. The production of an e.m.f. by heating the junction of two different metals (a *thermocouple*) is of only minor importance, but magnetism is the basis of all electrical generating machinery, and chemical changes are very useful on a smaller scale in batteries.

2.16 Instruments for Measuring Electricity

All three effects can be used for indicating the strength of currents; but although thermocouple ammeters (Sec 5.5) are used in radio

The same milliammeter can of course be used for measuring currents up to 1mA by connecting it in series in the current path; but to avoid increasing the resistance of the circuit more than need be, the voltage resistances would have to be short-circuited or otherwise cut out. Care must be taken never to connect a current meter directly across a voltage source, for its low resistance might pass sufficient current to destroy it. Higher currents can be measured by diverting all but a known fraction of the current through a bypass resistance called a *shunt*. If, for example, the resistance of the 0—1 milliammeter is 75Ω, shunting it with 75/9 = 8.33Ω would cause nine times as much current to flow through the shunt as passes through the meter, thereby multiplying the range by 10 (Fig. 2.15*b*).

Thus a single moving-coil instrument can be made to cover many ranges of current and voltage measurement, simply by connecting known resistances as shunts and multipliers. The operation can be reversed by incorporating a battery sufficient to give full-scale deflection; then, when an *unknown* resistance is connected, either as a shunt or multiplier, according to whether it is small or large, the extent to which the reading is reduced depends on the value of that resistance. It is therefore possible to graduate the scale in ohms and we have an *ohmmeter* (Fig 2.15*c*).

2.17 Electrical Power

When everyday words have been adapted for scientific purposes by giving them precise and restricted meanings, they are more likely to be misunderstood than words that were specially invented for use as scientific terms—'electron', for example. *Power* is a word of the first kind. It is commonly used to mean force, or ability, or authority. In technical language it has only one meaning: rate of doing work. *Work* here is also a technical term, confined to purely physical activity such as lifting weights or forcing things to move against pressure or friction—or potential difference. If by exerting muscular force you lift a 10 kg weight 0.5 m against the opposing force of gravity you do 10 × 0.5 = 5 kgm of work. And if you take 1 second to do it, your rate of doing work—your power output—is 5kg m per second (which is about (1/20th of a horse-power). Correspondingly, if the electromotive force of a battery raises the potential of 10 electrons by 3V (i.e., causes them to flow against an opposing p.d. of 3V) it does 30 electron-volts (eV) of work; and if it takes one second to do it the output of power is 30 volt-electrons per second. The electron per second is, as we saw, an extremely small unit of electric current, and it is customary to use one more than 6×10^{18} times larger—the ampere. So the natural choice for the electrical unit of power might be called the volt-ampere. Actually, for brevity (and for another reason which appears in Chapter 6), it has been given the name *watt*.

current can only flow from + to –, because a resistance is essentially a user-up of power.

2.18 A Broader View of Resistance

It will be found later on in this book that there are other ways in which electrical power can be 'lost' than by heating up the circuit through which the current flows. This leads to a rather broader view of resistance; instead of defining it as E/I, resulting from an experiment such as that connected with Fig. 2.3, one defines it as P/I^2 (derived from $P = I^2R$), P being the expenditure of power in the part of the circuit concerned, and I the current through it. The snag is that P is often difficult to measure. But as a definition to cover the various sorts of 'resistance' encountered in radio it is very useful.

A feature of electrical power expended in resistance—however defined—is that it leaves the circuit for good. Except by some indirect method it is not possible to recover any of it in the form of electrical power. But there are ways in which electrical power can be employed to create a store of energy that can be released again as electrical energy. *Energy* here is yet another technical term, meaning the amount of work that such a store can do. It can be reckoned in the same units as work, and so power can be defined alternatively as rate of release of energy. To return to our mechanical analogy, 5 kg m of energy expended by exerting a force of 10kg to push a box 0.5m across the floor is all dissipated as heat caused by friction.* But 5kg m used in raising a 10kg weight 0.5m is stored as what is called potential energy. If the weight is allowed to descend, it delivers up the 5kg m of energy, which could be used to drive a grandfather clock for a week. Another way of storing the energy would be to push a heavy truck mounted on ball bearings along rails. In this case very little of the pressure would be needed to overcome the small amount of friction; most of it would have the effect of giving it momentum which would keep it going long after the push had finished. A heavy truck in motion is capable of doing work because the force setting it in motion has been used to store energy in it; this kind is called kinetic energy.

In the next chapter we shall begin the story of how electrical energy can be stored and released in two ways, corresponding to potential and kinetic energy, and how this makes it possible to tune in to different stations on the radio or television.

---

*Electrical units are defined in terms of mechanical forces, etc., in metric units. Because of the relationships between metric and British units, a power of 746 watts is equal to 550 ft-lb per second (1 horse-power). It is confirmed by experiment that the rate of heating is the same, whether produced by dissipation of 1 mechanical HP or 746 electrical watts.

CHAPTER 3

Capacitance

3.1 Charging Currents

Although most of the last chapter was about steady e.m.fs driving currents through resistances, you may remember that the whole thing began with the formation of an electric charge. We likened it to the deportation of unwilling citizens from town A to town B. Transferring a quantity of electrons from one place to another and leaving them there sets up a stress between the two places, which is only relieved when the same quantity of electrons has been allowed to flow back. The total amount of the stress is the p.d., measured in volts. And the place with the surplus of electrons is said to be negatively charged. The original method of charging—by friction—is inconvenient for most purposes and has been generally superseded by other methods. If an e.m.f. of, say, 100 V is used, electrons will flow until the p.d. builds up to 100V; then the flow will cease because the charging and discharging forces are exactly equal.

It was at this stage that we provided a conductive path between the negatively and positively charged bodies, allowing the electrons to flow back at a rate proportional to the p.d.; and, as this p.d. is being maintained by a constant e.m.f., a continuous steady current is set up. This state of affairs could be represented as in Fig. 3.1. Our attention then became attached to this continuous current and the circuit through which it flowed. The electron surplus at the negative end—and the deficit at the positive end—seemed to have no bearing on this, and dropped out of the picture. So long as e.m.f. and current are quite steady, the charges can be ignored.

But not during the process of charging. While that is going on there must be more electrons entering B than are leaving it; and vice versa during discharging. Assuming that there are parts of a circuit requiring a substantial number of electrons to charge them to a potential equal to the applied e.m.f., then it is clear that the ordinary circuit principles we have been studying become more complicated whenever the e.m.f. varies. For one thing, it seems as if the current is no longer the same in all parts of a series circuit. The moment the e.m.f. is applied, a charging current starts to flow, over and above any current through conducting paths. If the e.m.f. is increased at any time, a further charging current flows to raise the p.d. correspondingly. Reducing the e.m.f. causes a discharging current. Since telecommunications and most other applications of electronics involve e.m.fs that are continually varying, it is evidently important to know how far the charging and discharging currents affect matters.

3.2 Capacitance: What It Is

The first thing is to find what decides the number of electrons needed to charge any part of a circuit. The number of people that could be transferred to another town would probably depend on (*a*) the pressure brought to bear on them, (*b*) the size of the town and (*c*) its remoteness. The number of electrons that are transferred as a charge certainly depends on the electrical pressure applied—the e.m.f. As one would expect, it is proportional to it. If the battery in Fig. 3.1 gave 200 V instead of 100, the surplus of electrons forced into B would be just twice as great. The quantity of electrons required to charge any part of a circuit to 1 V is called its *capacitance.*

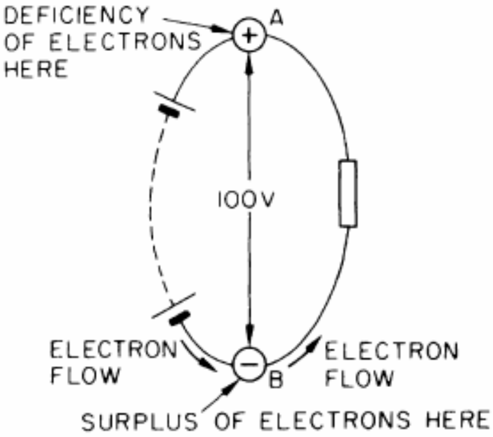


Fig. 3.1 Before a p.d. can be established between any two parts of a circuit (such as A and B) they must be charged, and when the p.d. is rapidly varied the charging current may be important

We have already noted in Sec. 2.5 that as a practical unit of electric current one electron-per-second would be ridiculously small and the number of electrons per second that fits the metric system (SI) is rather more than 6×10^{18} (= 1 A). So it is natural to take the same number of electrons as the unit of electric charge or *quantity* of electricity (symbol: Q). It is therefore equal to the number that passes a given point every second when a current of 1 A is flowing, and the name for it is the *coulomb.* So the unit of capacitance is that which requires 1 coulomb to charge it to 1 volt. It is named the *farad* (abbreviation: F) in honour of Michael Faraday, who contributed so much to electrical science. In symbols, the relationship between electric charge, voltage and capacitance is

$$Q = VC$$

which should be compared with the relationship between electric current, voltage, and conductance (Sec. 2.12). V is used instead of E because fundamentally it is the p.d. that is involved rather than the charging e.m.f. If the conducting path and e.m.f. in Fig. 3.1 were both removed, leaving A and B well insulated, their charge and the p.d. would remain. In practice one need not bother to distinguish between e.m.f. and p.d. every time (Sec. 2.14). All that matters is the voltage, and E and V do equally well.

As capacitance is equal to the charge required to set up a given potential *difference* between two parts of a circuit or other objects, one has to speak of the capacitance *between* those objects, or of one to another.

3.3 Capacitance Analysed

We saw that when any object is being charged the current going into it is greater than that coming out. So if Kirchhoff's law about currents is true in all circumstances, the object must be more than just a point; it must have some size. We would expect the capacitance of objects to one another to depend in some way on their size; the question is, what way?

When analysing resistance (Sec. 2.11) we found that it depended on the dimensions of the conductor and what it was made of. So, of course, does conductance, though in inverse ratio (Sec. 2.12). Capacitance, however, has to do with the space between the conductors; it results from the force of attraction that can be set up in that space, or from what is called the electric field. Suppose A and B in Fig. 3.2 are our oppositely charged terminals. The directions along which the force acts are indicated by dashed lines, called *lines of electric force*. These show the paths electrons would take if they were free to move slowly under the force of attraction to +. (But to agree with the convention for direction of electric currents the arrows point in the opposite direction). To represent a more intense field, lines are drawn more closely together. Each can be imagined to start from a unit of positive charge and end on a unit of negative charge. It must be understood that it is a matter of convenience what size of unit is taken for this purpose; and that the lines are not meant

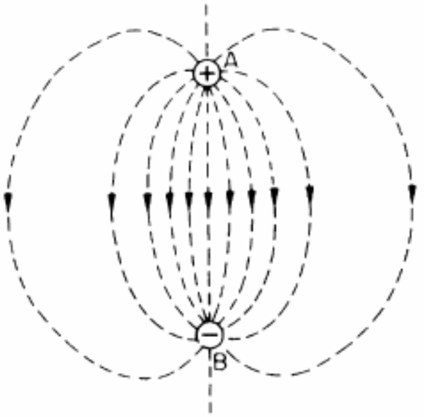


Fig. 3.2 The dashed lines indicate the direction and relative intensity of electric field between two oppositely charged terminals, A and B

3.4 Capacitors

The general symbol for a capacitor in circuit diagrams is Fig. 3.4*a*. The capacitors themselves appear in great variety in circuits, according to the required capacitance and other circumstances. Some actually do consist of a single sheet of insulating material sandwiched between a pair of metal plates, but sometimes there are a number of plates on each side as indicated in Fig. 3.4*b*. The effective area of one dielectric is of course multiplied by their number—in this case four.

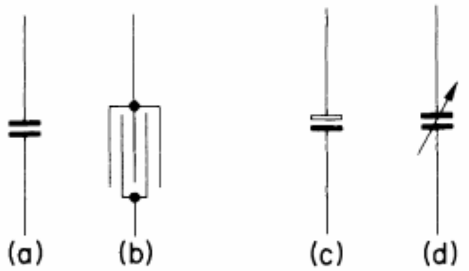


Fig. 3.4 The standard symbol for a fixed capacitor is at a: it often has many plates interleaved as at b. Electrolytic capacitors are distinguished as at c, in which the white plate is positive. d denotes a variable capacitor

A very common form of capacitor consists of two long strips of aluminium foil separated by polyester film and rolled up into a compact block. Above a few microfarads it is often more economical to separate the aluminium electrodes by chemically impregnated paper or other material which causes an extremely thin insulating film to form on one of them. Apart from this film-forming, the chemical solution also acts as a conductor constituting one plate of the capacitor. The aluminium foil is the other plate and the two plates are separated by the film which acts as dielectric. These *electrolytic* capacitors must always be connected and used so that the terminal marked + never becomes negative relative to the other. Even when correctly used there is always a small leakage current. To help the wireman, the symbol Fig. 3.4*c*, in which the + plate is distinguished by outlining, is used for electrolytic types.

Air is seldom used as the dielectric for fixed capacitors, but is quite common in variable capacitors. These, indicated in diagrams as 3.4*d*, consist of two sets of metal vanes which can be progressively interleaved with one another, so increasing the effective area, to obtain any desired capacitance up to the maximum available, seldom more than 500pF. Plastic film dielectric is used in variable capacitors where compactness is essential.

Besides the capacitance, an important thing to know about a capacitor is its maximum working voltage. In Sec. 2.3 we visualised insulators as substances in which the electrons are unable to drift through the material in large numbers under the influence of an electric field but instead they 'strain at the leash', being slightly displaced from their normal positions in the atoms. If the intensity of electric field is increased beyond a certain limit the elastic leashes tethering the electrons snap under the strain, and the freed electrons rush uncontrollably to the positive plate, just as if the dielectric were a good conductor. It has, in fact, been broken down or punctured by the excessive voltage. For a given thickness, mica stands an exceptionally high voltage, or, as one says, has a relatively high dielectric strength. Air is less good, but has the advantage that the sparking across resulting from breakdown does it no permanent harm.

In our concentration on intentional capacitance between plates, we should not forget that, whether we like it or not, every part of a circuit has capacitance to surrounding parts. When the circuit e.m.fs are varying rapidly, these stray capacitances may be very important.

Semi-conductor diodes can also behave as capacitors and these are described in Chapter 9.

3.5 Charge and Discharge of a Capacitor

It is instructive to consider the charging process in greater detail. In Fig. 3.5 the capacitor can be charged by moving the switch to A, connecting it across a battery; and discharged by switching to B. R might represent the resistance of the wires and capacitor plates, but as that is generally very small it will be easier to discuss the matter if we assume we have added some resistance, enough to bring the total up to, say, 200Ω. Suppose the capacitance C is 5μF and the battery e.m.f. 100V.

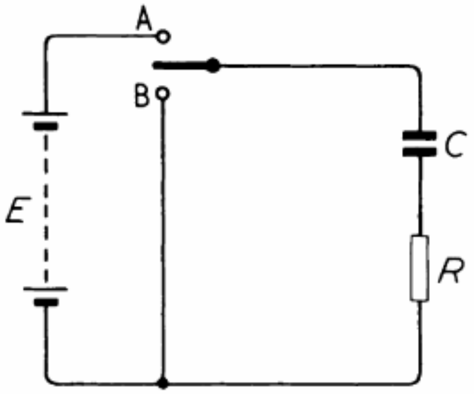


Fig. 3.5 Circuit used for obtaining the charge and discharge curves shown in Fig. 3.6

At the exact moment of switching to A the capacitor is as yet through R, and so the current becomes less and the capacitor charges more slowly. When it is charged to 50V, 50V are left for the resistor, and the charging rate is 0.25 A. When C is charged to 80V, 20 are left for R, and the current is 0.1A. And so on, as shown by the curves in Fig. 3.6.

Curves of this type are known as *exponential*, and are characteristic of many growth and decay phenomena. Note that the current curve's downward slope, showing the rate at which the current is decreasing, is at every point proportional to the current itself. It can be shown mathematically that in the time a capacitor would take to charge at the starting rate it actually reaches 63.2% of its full charge. By using the principles we already know we have calculated that this time in the present example is 0.001 s. Try working it out with letter symbols instead of numbers, so that it covers all charging circuits, and you should find that it is CR s, regardless of the voltage E. CR—equal to the time in seconds taken to reach 63.2% of final charge—is of course a characteristic of the circuit, and is called its *time constant.* C in this example being 5×10^{-6} F and R 200Ω, CR is 0.001s, confirming Fig. 3.6.

Theoretically, the capacitor is never completely charged, because the charging depends on there being a difference between the applied e.m.f. and the capacitor voltage, to drive current through the resistance. If the current scale is multiplied by 200 (the value of R), the upper curve is a graph of this voltage across R. Note that the voltages across C and R at all times up to 0.004s add up to 100, just as Kirchhoff's voltage law says they ought (Sec. 2.13).

For practical purposes the capacitor may be as good as fully charged in a very short time. Having allowed 0.004 s for this to happen in the present case, we move the switch to B. Here the applied voltage is zero, so the voltages across C and R must now add up to zero. We know that the capacitor voltage is practically 100, so the voltage across R must be -100, and the current -0.5A; that is to say, it is flowing in the opposite direction, discharging the capacitor. And so we get the curves in Fig. 3.6 from 0.004s onwards.

A point to remember is that the voltages set up across R at the instants of switching to A and B are short-lived; some idea of their duration in any particular case can be found by multiplying C by R in farads and ohms respectively, giving the answer in seconds. These temporary changes are called *transients.*

3.6 Where the Power Goes

The height of the charging curve above the zero line in Fig. 3.6*a* represents the current supplied by the battery. Multiplied by its e.m.f. (100V) it represents the power. When first switched on, the battery is working at the rate of 0.5 × 100 = 50W, but at the end of

0.004s this has fallen off almost to nothing. At first the whole 50W goes into the resistance, so all its energy is lost as heat. But the effort of the e.m.f. immediately becomes divided, part being required to drive the current through *R* and the rest to charge *C*. It is clear that energy is going into the capacitance (Sec. 2.18); what happens to it? The answer is that it is stored as an electric field, in a similar way to the storage of mechanical energy in a spring when it is compressed by an applied force. During the charging process as a whole, then, part of the release of energy from the battery is lost as heat in the resistor, and part is stored in the capacitor.

When we come to the discharge we find that the voltage across *R* and the current through *R* follow exactly the same curves as during the charge, except that they are both reversed. But multiplying two negatives together gives a positive, so the energy expended in *R* is the same amount as during charge. The capacitor voltage and current during the discharge are the same as for the resistor, so the amount of energy is equal; in fact, it is the same energy, for what goes into the resistor has to come out of the capacitor. Applying the test described at the end of Sec. 2.17, we can check that the capacitor is a source of energy; the current that is going into the resistor at its positive terminal is coming out from the capacitor's positive terminal.

At the end of the discharge, all the energy supplied by the battery during the charging process has been dissipated in the resistance as heat. And as the amounts dissipated therein during charge and discharge are equal, it follows that during charge the energy from the battery was shared equally by *R* and *C*.

Summing up: during the charging period half the energy supplied by the battery is stored as an electric field in the capacitor and half is dissipated in the resistor; during discharge the half that was stored is given back at the expense of the collapsing field and is dissipated in the resistor.

It may be as well to repeat the warning that the 'charging' and 'discharging' of accumulator batteries is an entirely different process, accompanied by only minor variations in voltage. The energy put into the battery is not stored as an electric charge; it causes a reversible chemical change in the plates.

3.7 Recapitulation

By this time you may be getting a little confused about the meanings and relationships of the various quantities we have met. So let us recapitulate. First, those that do not necessarily depend on time. There is electric charge (*Q*), reckoned in coulombs (C), which can be thought of as surpluses or deficiencies of electrons. Transferring electrons from A to B charges A positively and B negatively (as we say), and set up a p.d. (*V*) between them, which is a kind of stress and is reckoned in volts (V). Q/V is called the capacitance (C) between A and B, and its unit is the farad (F). To make the electrons move from A to B against the p.d. created by the separation of the charges we need an e.m.f. (E) which is reckoned in terms of the p.d. it is able to counteract, so its unit also is the volt. The p.d. is always tending to restore the charge to the normal uncharged state, and if it is allowed to do so the electrical energy stored in C is released. That energy is equal to the effort (called work) that the source of e.m.f. had to do in order to charge C. Because energy and work, like income and expenditure, are the same thing looked at from different points of view, they are both denoted by W and have the same unit—the *joule* (J)—new information, this.

We know (Sec. 3.2) that if a charge Q is imparted to a capacitance C the resulting p.d., V, is equal to Q/C. But as the starting p.d. was zero and increased in proportion to the charge, the *average* p.d.—and therefore the average e.m.f. needed to transfer the charge—must be half as much, $Q/2C$. The work done by the e.m.f., and therefore the energy stored in C, is Q multiplied by the average e.m.f., that is to say

$$W = Q \times Q/2C = Q^2/2C = V^2C^2/2C = \tfrac{1}{2}V^2C.$$

The important thing to remember about energy is that it never ceases to exist. Even when the switch in Fig. 3.5 is moved to B the electrical energy stored in C is converted into an equal amount of heat energy in R.

Time comes into things as rates of happening. The rate at which charge is moved is called current (I), reckoned in amperes (A). So if a current of I amps continues steadily for ts the charge moved past a fixed point en route is It coulombs and $I = Q/t$. If the charge is moved against a constant p.d. V, energy equal to VQ is transferred every second, and an equal amount of work has to be done by a source of e.m.f. The rate at which energy is transferred and work is done is called power (P) and is reckoned in watts (W). So $P = VQ/t = VI$ (Sec. 2.17).

3.8 Displacement Currents

If any readers are worrying about reconciling capacitance with Kirchhoff's current law, which implies that current in a series circuit is the same throughout, they deserve a note on the subject.

There should be no difficulty so far as the extra capacitance due to dielectrics is concerned. A charging or discharging current in the circuit connected to the capacitor plates can be regarded as continued through the dielectric by the displacements of electrons as they 'strain at the leash'. The fact that this movement is very restricted is no problem, for it is consistent with the very limited

CHAPTER 4

Inductance

4.1 Magnets and Electromagnets

If a piece of paper is laid on a straight bar magnet, and iron filings are sprinkled on the paper, they are seen to arrange themselves in a pattern something like Fig. 4.1*a*. Gently tapping the paper helps the filings to take up the pattern. The lines show the paths along which the attraction of the magnet exerts itself. (Compare the lines of electric force in Fig. 3.2.) As a whole, they map out the *magnetic field,* which is the sphere of influence, as it were, of the magnet. The field is most concentrated around two regions, called *poles,** at the ends of the bar. The lines may be supposed to continue right through the magnet, emerging at the end marked N and returning at S. This direction, indicated by the arrows, is (like the direction of an electric current) purely conventional (representing the direction in which a North pole would be moved), and the lines themselves are an imaginary representation of a condition occupying the whole space around the magnet.

The same result can be obtained with a previously quite ordinary and unmagnetised piece of iron by passing an electric current through a wire coil round the iron — an arrangement known as an *electromagnet.* It is not even necessary to have the iron core (as it is called); the coil alone, so long as it is carrying current, is interlinked with a magnetic field having the same general pattern as that due to a bar magnet, as can be seen by comparing Fig. 4.1*a* and *b*. Without the iron core, however, it is considerably weaker. Finally, if the wire is unwrapped, every inch of it is still found to be surrounded by a magnetic field. Though very much less concentrated than in the coil, if the current is strong enough it can be demonstrated with filings as in Fig. 4.2.

The results of these and other experiments in electromagnetism are expressed by saying that wherever an electric current flows it surrounds itself with a field which sets up magnetic *flux* (symbol: Φ, the Greek capital phi). This flux and the circuit link one another like two adjoining links in a chain. By winding the circuit into a compact coil, the field due to each turn of wire is made to operate in the same space, producing a concentrated magnetic flux (Fig. 4.1*b*). Finally,

---

*The terminals of a battery or other source of e.m.f., between which electric field lines are imagined, are sometimes also called poles. The distinction between + and – in such sources, and N and S in magnets, is called its *polarity.*

4.2 Interacting Magnetic Fields

The 'N' and 'S' in Fig. 4.1 stand for North-seeking and South-seeking respectively. Everybody knows that a compass needle points to the north. The needle is a magnet, and turns because its own field interacts with the field of the earth, which is another magnet. Put differently, the north magnetic pole of the earth attracts the north-seeking pole of the needle, while the earth's south pole attracts its south-seeking pole. The poles of a magnet are often called just north and south, but strictly, except in referring to the earth itself, this is incorrect. By bringing the two poles of a bar magnet — previously identified by suspending it as a compass — in turn towards a compass needle it is very easy to demonstrate, as in Fig. 4.3, that unlike poles attract one another. This reminds one of the way electric charges behave.

Now suppose we hang a coil in such a way that it is free to turn, as suggested in Fig. 4.4. So long as no current is passing through the coil it will show no tendency to set itself in any particular direction, but if a battery is connected to it the flow of current will transform the coil into a magnet. Like the compass needle, it will then indicate the north, turning itself so that the plane in which the turns of the coil lie is east and west, the axis of the coil pointing north. If now the current is reversed the coil will turn through 180 degrees, showing that what was the N pole of the coil is now, with the current flowing the opposite way, the S.

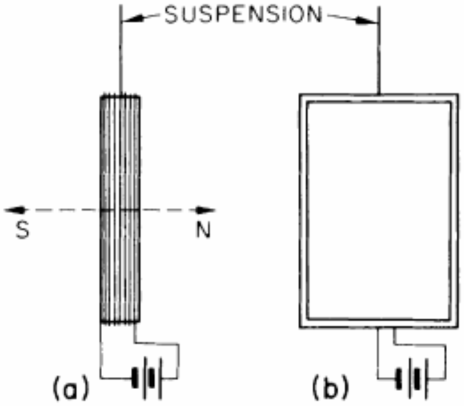


Fig. 4.4 A coil carrying current, and free to rotate, sets itself with its axis pointing N and S, (*a*) is a side view and (*b*) an end view

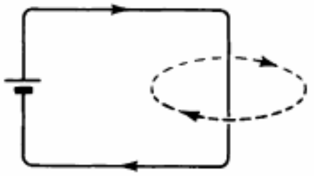
The earth's field is weak, so the force operating to turn the coil is small. When it is desired to make mechanical use of the magnetic effect of the current in a coil it is better to provide an artificial field of the greatest possible intensity by placing a powerful magnet as close to the coil as possible. This is the basis of all electric motors.

In electronics, however, we are more interested in other applications of the same principle, such as the moving-coil measuring instruments described very briefly in Sec. 2.16. The coil is shaped somewhat as in Fig. 4.4, and is suspended in the intense field of a permanent magnet. Hairsprings or taut metal ribbons at top and bottom of the coil serve the double purpose of conducting the current to the coil and keeping the coil normally in the position where the pointer attached to it indicates zero. When current flows through it the coil tends to rotate against the restraint of the springs, and the angle through which it moves is proportional to the magnetic interaction, which in turn is proportional to the value of the current.

4.3 Induction

In Sec. 4.1 we noted that the magnetic circuit formed by the imaginary flux lines is linked with the electric circuit that causes them, like two adjoining links in a chain (Fig. 4.5). Now suppose the electric circuit includes no source of e.m.f. and consequently no current, but that some magnetic flux is made to link with it. There are various ways in which this can be done: a permanent magnet or an electromagnet—or any current-carrying circuit—could be moved towards it, or current could be switched on in a stationary circuit close to it. Whatever the method, the result will be the same: so long as the flux linked with the circuit is changing, an e.m.f. is generated in the circuit. The magnitude and direction of this *induced* e.m.f. can be found by means of a suitable voltmeter, as in Fig. 4.6.

Fig. 4.5 An electric current and the magnetic flux set up thereby are linked with one another as indicated in this diagram, which shows the relative directions established by convention



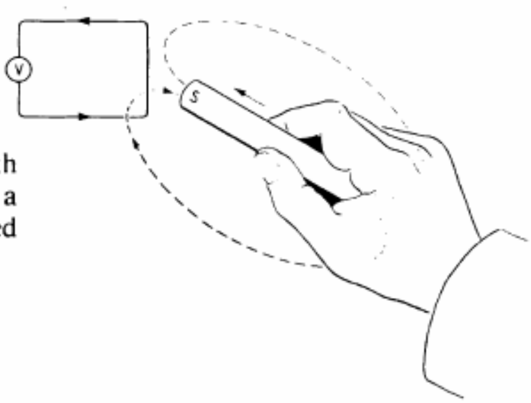


Fig. 4.6 When the magnetic linkage with any closed path is varied, say by moving a magnet near it, an e.m.f. is generated around the path

The amount of e.m.f. is proportional to the rate at which the amount of flux linked with the circuit is changing. So a constant flux linkage yields no e.m.f. When the flux is removed or decreased, a reverse e.m.f. is induced. In Fig. 4.6 the increasingly linked flux has the same direction as in Fig. 4.5; the e.m.f. induced thereby would tend to drive current in the direction shown; i.e., in the opposite direction to the current in Fig. 4.5.

These results, first discovered by Michael Faraday in 1831, are the basis of all electrical engineering, including of course radio. The electrical generators in power stations are devices for continuously varying the magnetic flux linked with a circuit, usually by making electromagnets rotate past fixed coils. The alternative method of varying the flux linkages—by varying the current in adjacent fixed coils—is adopted in transformers (Sec. 4.11) and, as we shall see, in many kinds of radio equipment.

The SI unit of flux, the *weber* (Wb), is such that the induced e.m.f. in volts is equal to the rate at which the linked flux is changing, in webers per second.

4.4 Self-Inductance

We have just been considering an e.m.f. being induced in a circuit by the varying of magnetic flux due to current in some other circuit. But in Fig. 4.5 we see a dashed loop representing the flux due to current flowing in that circuit itself. Before the current flowed there was no flux. So, directly the current was switched on, the flux linked with the circuit must have been increasing from zero to its present value, and in the process it must have been inducing an e.m.f. in its own circuit. This effect is—very naturally—known as *self-induction.* Since Fig. 4.5 can be taken to represent *any* circuit carrying current, and because a flow of current inevitably results in some linked magnetic flux, it follows that an e.m.f. is induced in any circuit in which the current is varying.

The amount of e.m.f. induced in a circuit (or part of a circuit) when the current is varied at some standard rate is called its *self-inductance,* or often just inductance (symbol: L). Its unit, the *henry* (abbreviation: H), was so defined that the inductance of a circuit in henries is equal to the number of volts self-induced in it when the current is changing at the rate of 1 ampere per second.

Actually it is the change of *flux* that induces the e.m.f., but since information about the flux is less likely to be available than that about the current, inductance is defined in terms of current. What it depends on, then, is the amount of flux produced by a given current, say 1 A. The more flux, the greater the e.m.f. induced when that current takes one second to grow or die, and so the greater the inductance. The flux produced by a given current depends, as we have seen, on the dimensions and arrangement of the circuit or part of circuit, and on the permeability.

4.5 Lenz's Law

In Sec 4.3 we noted the direction of the induced e.m.f. in relation to the direction of the flux inducing it. By comparing Fig. 4.6 with Fig. 4.5 we see that a self-induced e.m.f. would tend to send current round the circuit in the opposite direction to the current producing the flux that is inducing it. In more general terms, *any* induced e.m.f. tends to oppose its cause. This statement is known as Lenz's law, and is really inevitable, because if induced e.m.fs aided their cause there would be no limit to their growth. So it is easy to remember.

When you try to stop the clockwise current in Fig. 4.5 by switching off, the flux in the direction shown begins to *decrease,* and in doing so it induces an e.m.f. in the *same* direction as that of the battery, tending to oppose the switch-off and keep the current going.

This is a good point at which to learn another easily remembered rule—the corkscrew rule. Imagine that the current in Fig. 4.2 is a corkscrew being driven downwards. The direction the corkscrew has to be turned is the direction of the resulting flux. This can be seen in Fig. 4.5 too. And it holds good if the roles are interchanged; turning the corkscrew in the direction of the current around the circuit, the direction in which it moves, in or out, is the direction of the resulting flux through the circuit. With this rule and Lenz's law we are fully equipped to deal with induction polarity questions. But we must realise that the directions are conventions, not physical facts.

4.6 Inductance Analysed

We had no great difficulty in finding a formula expressing resistance in terms of resistivity and the dimensions of the circuit element concerned, because currents are usually confined to wires or other conductors of uniform cross-section. Capacitance was not so easy, because it depends on the dimensions of space between conductors, and the distribution of the electric field therein is likely to be anything but uniform, but for the particular and practically-important case of a thin space between parallel plates we found a formula for capacitance in terms of permittivity and the dimensions of the space. Inductance is even more difficult, because the magnetic fields of many of the circuit forms used for providing it, known in general as *inductors,* are not even approximately uniform or easily calculated. And some of the flux may link with only part of the circuit. In coils, for example, the flux generated by one turn usually links with some of the others but not all.

However, Fig. 4.7 shows a type of inductor in which the coil is linked with an iron core of nearly uniform cross-section area. The permeability of the iron is usually so high that the core carries very nearly all the magnetic flux; the remainder—called *leakage* flux—

4.7 Practical Considerations

In practice this formula is not of much use for calculating the inductance of even iron-cored coils. For one thing, in order to get the core into position around the coil it is necessary to have it in more than one piece, and the joint is never so magnetically perfect as continuous metal. Even a microscopic air gap makes an appreciable difference, because it is equivalent to an iron path perhaps tens of thousands of times as long. Then the permeability of iron depends very much on the flux density and therefore on the current in the coil. So one cannot look it up in a list of materials as one can resistivity or permittivity. It is usually plotted as a graph against the strength of the magnetising field, which has the symbol H and is reckoned as the total current encircling the iron (i.e., the number of ampere-turns) per unit length of iron carrying the flux. For many purposes it is even more helpful to have graphs of the flux density per square metre, which has the symbol B and is reckoned in teslas (i.e., webers of flux per square metre) against H, as in Fig. 4.8. The permeability, equal to B/H, can be found from this for any particular value of H or B. It is represented by the steepness of the slope of the graph. A characteristic of ferromagnetic materials is that increasing the ampere-turns yields progressively smaller increases in flux density. In other words, the permeability gets less and less. This effect is called *magnetic saturation.*

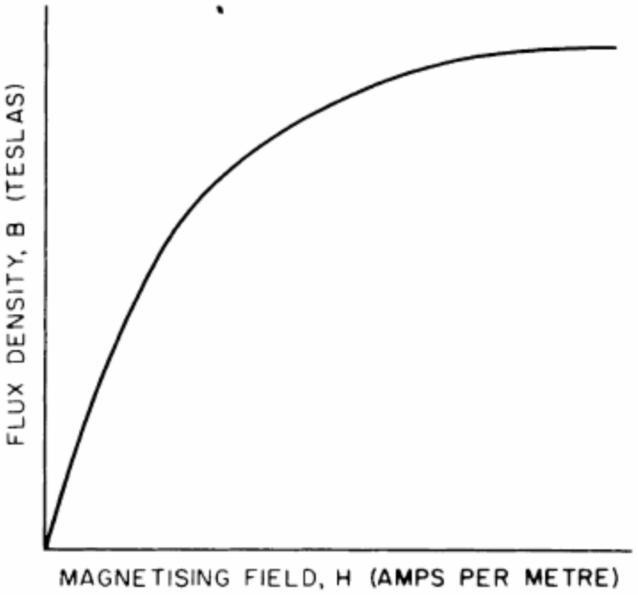


Fig. 4.8 The flux density in many magnetic materials such as iron varies with the magnetising field in this sort of way.

If you are finding this section rather heavy going, with so many new quantities cropping up, comfort yourself with the assurance that they will not be found essential within the scope of this book.

But if you are keen enough to pursue the magnetic relationships farther you may care to note, starting from the fact that $\mu = B/H$, that one can derive the first formula for inductance in Sec. 4.6:

$$\mu = \frac{B}{H} = \frac{\Phi l}{AI}$$

So

$$\frac{\mu A}{l} = \frac{\Phi}{I} = L$$

With ‘air-core’ coils, even when all the turns are close together it is not correct to assume that all the flux due to each turn links with the whole lot; and the error is of course much greater if the turns are widely distributed as in Fig. 4.1*b*. An even greater difficulty is that the flux does not all follow a path of constant area A and equal length l.

A general formula for inductance has nevertheless been given, to bring out the basic similarity in form to those for conductance and capacitance, but for practical purposes it is necessary either to measure inductance or calculate it from the various formulae and graphs that have been worked out for coils of various shapes and straight wires and published in radio reference books.

Between a short length of wire and an iron-cored coil with thousands of turns there is an enormously large range of inductance. The iron-cored coil might have an inductance of many henries; the short wire a small fraction of a microhenry. For most purposes the inductance of connecting wires is small enough to ignore. But not always. The maximum current may be much less than 1 A, but if it comes and goes a thousand million times a second its rate of change may well be millions of amps per second and the induced voltage quite large.

It is, of course, this induced voltage that makes inductance significant. Just how significant will appear in later chapters, but we can make a start now by comparing and contrasting it with the voltage that builds up across a capacitance while it is being charged.

4.8 Growth of Current in an Inductive Circuit

What happens when current is switched on in an inductive circuit can be studied with the arrangement shown in Fig. 4.9, which should be compared with Fig. 3.5. To make comparison of the results easy, let us assume that the inductance, for which the scallops marked L is the standard symbol, is 0.2 H, and that R (which includes the resistance of the coil) is 200 Ω and E is 100 V as before.

With the switch as shown, no current is flowing through L and there is no magnetic field. At the instant of switching to A, the full 100 V is applied across L and R. The current cannot instantly leap to 0.5 A, the amount predicted by Ohm's law, for that would mean

R. And as the applied voltage remains 100, the induced voltage must diminish. The only way this can happen is for the current to grow less rapidly. By the time it has reached 0.25 A there are 50 V across *R*, therefore only 50 across *L*, so the current must be growing at half the rate. The full line shows how it grows; here is another exponential curve and, as in the capacitive circuit, it theoretically never quite finishes. In the time that the current would take to reach its full Ohm's law value if it kept up its starting pace, it actually reaches 63.2% of it. This time is, as before, called the time constant of the circuit. A little elementary algebra based on the foregoing shows it to be *L*/*R* seconds.

The voltage across *L* is also shown. Compared with Fig. 3.6, current and voltage have changed places. The voltage across *R* is of course proportional to the current and therefore its curve is the same shape as Fig. 4.10*b*. Added to the upper, it must equal 100 V as long as that is the voltage applied.

When the current has reached practically its full value, we flick the switch instantaneously across to B. The low resistance across the contacts is merely to prevent the circuit from being completely interrupted in this process. At the moment the switch is operated, the magnetic field is still in existence. It can only cease when the current stops, and the moment it begins to do so an e.m.f. is induced which tends to keep it going. At first the full 0.5 A is going, which requires 100 V to drive it through *R*; and as the battery is no longer in circuit this voltage must come from *L*, by the current falling at the required rate: 500 A/s. As the current wanes, so does the voltage across *R*, and so must the induced voltage, and therefore the current dies away more slowly, as shown in the continuation of Fig. 4.10. In *L*/*R* (= 0.001 s) it has been reduced by the inevitable 63.2%.

4.9 Power During Growth

In reckoning how the power in the circuit varies during these operations, we note firstly that the output from the battery (calculated by multiplying the current at each instant by its 100 V) instead of starting off at 50 W and then tailing off almost to nothing, as in the capacitive example, starts at nothing and works up towards 50 W. If there had been no *C* in Fig. 3.5 there would have been no flow of power at any time; the effect of *C* was to make possible a temporary flow, during which half the energy was dissipated in *R* and half stored in *C*. In Fig. 4.9, on the other hand, absence of *L* would have meant power going into *R* at the full 50-W rate all the time: the effect of *L* was to withhold part of that expenditure from *R* for a short time. Not all this energy withheld was a sheer saving to the battery, represented by the temporarily sub-normal current; part of it—actually a half—went into *L*, for the current had to be forced into motion through it against the voltage induced by the growing magnetic field. The energy was in fact stored in the magnetic field, in a similar way to the storage of mechanical energy when force is applied to set a heavy vehicle in motion.

Just as the vehicle gives up its stored energy when it is brought to a standstill, the result being the heating of the brakes, so the magnetic field in L gave up its energy when the current was brought to a standstill, the result being the heating of R. Self-inductance is therefore analogous to mechanical inertia.

The amount of energy stored when a current I is flowing is again equal to the charge Q transferred during the magnetising process, multiplied by the p.d. it has been transferred against. Supposing the current grew at the starting rate in Fig. 4.10 it would reach its full value I in time L/R, so $Q = IL/R$. The average p.d. across the coil is clearly half the maximum value IR. So

$$W = \frac{IL}{R} \times \tfrac{1}{2}IR = \tfrac{1}{2}I^2L$$

If we write this in the form $\frac{1}{2}LI^2$ it compares directly with $\frac{1}{2}mv^2$, the expression for kinetic energy. Just as $\frac{1}{2}LI^2$ expresses the energy stored in the magnetic field of an inductor, so $\frac{1}{2}mv^2$ gives the energy stored in a body of mass m moving at a velocity v. Thus inductance is analogous to mass and current to velocity.

4.10 More Comparison and Contrast

In the previous chapter we saw that besides the conducted currents, reckoned by dividing voltage by resistance, there is an extra current whenever the voltage is varying. This current is made up of movements of charges caused by the electric field adjusting itself to the new voltage. It is proportional not only to the rate at which the voltage is changing but also the capacitance (intentional or otherwise) of the part of the circuit considered. In this chapter we have seen that in addition to the resistive voltage, reckoned by multiplying current by resistance, there is an extra voltage whenever the current is varying. This voltage is caused by the magnetic field adjusting itself to the new current. It is proportional not only to the rate at which the current is changing but also to the inductance (intentional or otherwise) of the part of circuit considered.

4.11 Mutual Inductance

If a coil with the same number of turns as that in Fig. 4.9 were wound so closely around it that practically all the magnetic flux due to the original *primary winding* embraced also this *secondary winding,* then an equal voltage would at all times be induced in the secondary.

CHAPTER 5

Alternating Currents

5.1 Frequencies of Alternating Current

In our first chapter we saw that speaking and music are conveyed from one person to another by sound waves, which consist of rapid vibrations or alternations of air pressure. And that to transmit them over longer distances by telephone it is necessary for electric currents to copy these alternations. And further, that for transmitting them across space, by radio, it is necessary to use electric currents alternating still more rapidly. Of course we must never forget that sound waves and radio waves are quite different in nature, sound waves being mechanical vibrations of material substances (solid, liquid or gas) whereas radio waves are oscillations of electric and magnetic fields which can cross interstellar space.

We have now just noted the fact that the great advantage of being able to step voltages up and down as required is obtainable only when the electricity supply is continually varying, which is most conveniently done by arranging for it to be alternating. Public electricity supplies which light and heat our houses and provide the power that works our television receivers and many other facilities are therefore mostly of the alternating-current kind.

The only essential difference between all these alternating currents is the number of double-reversals they make per second; in a word, the frequency. We have already gone fairly fully into the matter of frequency (Sec. 1.4), so there is no need to repeat it; but it may be worth recalling the main divisions of frequency.

There is no hard-and-fast dividing line between one lot of frequencies and another; but those below 100 Hz are used for power (the standard in Britain is 50 Hz); those between about 20 and 20 000 Hz are audible, and therefore are classed as audio frequencies (a.f.); while those above about 15 000 are more or less suitable for carrying signals across space, and are known as radio frequencies (r.f.). Certain of these, such as 525 000 to 1 605 000, are allocated for broadcasting. All those above about 1000 MHz (wavelengths shorter than about 30 cm) are often lumped together as *microwave* frequencies.

For convenience the enormous range of radio frequencies is divided into subranges (decades), each with a 10:1 ratio of maximum to minimum frequency (and wavelength), as shown in the accompanying table which also gives the terminology and the abbreviations in common use for the subranges.

Frequency range	*Wavelength range*	*Name*	*Abbreviation*
less than 30 kHz	greater than 10000 m	very low frequencies	v.l.f.
30 kHz–300 kHz	1000 m–10000 m	low frequencies	l.f.
300 kHz–3 MHz	100 m–1000 m	medium frequencies	m.f.
3 MHz–30 MHz	10 m–100 m	high frequencies	h.f.
30 MHz–300 MHz	1 m–10 m	very high frequencies	v.h.f.
300 MHz–3 GHz	10 cm–1 m	ultra high frequencies	u.h.f.
3 GHz–30 GHz	1 cm–10 cm	super high frequencies	s.h.f.
30 GHz–300 GHz	1 mm–1 cm	extremely high frequencies	e.h.f.

What has been said about currents applies to voltages, too; it requires an alternating voltage to drive an alternating current.

The last two chapters have shown that when currents and voltages are varying there are two circuit effects—capacitance and inductance—that have to be taken into account as well as resistance. We shall therefore be quite right in expecting a.c. circuits to be a good deal more complicated and interesting than d.c. Before going on to consider these circuit effects there are one or two things to clear up about a.c. itself.

5.2 The Sine Wave

An alternating current might reverse abruptly or it might do so gradually and smoothly. In Fig. 3.6*a* we have an example of abrupt reversal. There is, in fact, an endless variety of ways in which current or voltage can vary with time, and when graphs are drawn showing how they do so we get a corresponding variety of wave shapes. The steepness of the wave at any point along the time scale shows the rapidity with which the current is changing at that time.

Fortunately for the study of alternating currents, all wave shapes, however complicated, can be regarded as built up of waves of one fundamental shape, called the *sine wave* (adjective: sinusoidal). Figure 1.2*c* shows examples of this. Note the smooth, regular alternations of the component sine waves, like the swinging of a pendulum. Waves of even such a spiky appearance as those in Fig. 3.6 can be analysed into a number of sine waves of different frequencies.

Most circuits used in radio and other systems of telecommunications consist basically of sources of alternating e.m.f. (often called generators) connected to combinations of resistance, inductance and/or capacitance (usually called loads). With practice one can reduce even very complicated-looking circuits to standard generator-and-load combinations.

We have just seen that although there is no end to the variety of waveforms that generators can produce, they are all combinations of simple sine waves. So it is enough to consider the sine-wave gnerator. Sources of more complicated waveforms can be imitated by combinations of sine-wave generators. Theoretically this dodge is always available, but in practice (as, for example, with square waves) there are sometimes easier alternatives, as we shall see in Chapter 25. Unless the contrary is clear, it is assumed from now on that a generator means a sine-wave generator.

5.3 Circuit with Resistance only

The simplest kind of a.c. circuit is one that can be represented as a generator supplying current to a purely resistive load (Fig. 5.1). We have already found (Sec. 2.6) how to calculate the current any given e.m.f. will drive through a resistance. Applying this method to an alternating e.m.f. involves nothing new. Suppose, for example, that the voltage at the peak of each wave is 200, and that its frequency is 50 Hz. Figure 5.2 shows a graph of this e.m.f. during rather more than one cycle. And suppose that R is 200 Ω. Then we can calculate the current at any point in the cycle by dividing the voltage at that point by 200, in accordance with Ohm's law. At the peak it is, of course, 200/200 = 1 A; half-way up the wave it is 0.5 A. Plotting a number of such points we get the current wave, shown dashed. It is identical in shape, though different in vertical scale.

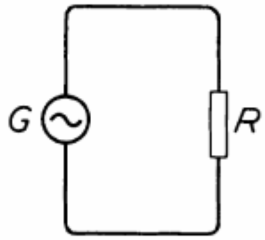


Fig. 5.1 Circuit consisting of an a.c. generator (*G*) feeding a purely resistive load (*R*)

5.4 R.M.S. Values

How is an electricity supply that behaves in this fashion to be rated? Can one fairly describe it as a 200 V supply, seeing that the actual voltage is changing all the time and is sometimes zero? Or should one take the average voltage?

The answer that has been agreed upon is based on comparison with d.c. supplies. It is obviously very convenient for a lamp or heater intended for a 200 V d.c. system to be equally suited to a.c. mains of the same nominal voltage. This condition will be fulfilled if the average *power* taken by the lamp or heater is the same with both types of supply, for then the element will reach the same temperature and the light or heat will be the same in both cases.

We know (Sec. 2.17) that the power in watts is equal to the voltage

Generalising this calculation, we find that when an alternating voltage is adjusted to deliver the same power to a given resistance as a direct voltage, its peak value is $\sqrt{2}$ times the direct voltage, or about 1.414 times as great. Put the other way round, its nominal or equivalent or effective voltage—called its *r.m.s.* (root-mean-square) value—is $1/\sqrt{2}$ or about 0.707 times its peak value. Since the resistance is the same in both cases, the same ratio exists between r.m.s. and peak values of the current. What is called a 240-V a.c. supply therefore alternates between + and $-\sqrt{2} \times 240 = 339$ V peak.

Since the r.m.s. values of voltage and current are 0.707 times the peak values, the power dissipated in a resistance is given by $0.707 \times E_{\max} \times 0.707 \times I_{\max} = \frac{1}{2} E_{\max} I_{\max}$, i.e. one half the product of the peak values. This is important because, when we come to the output power of certain amplifier types (Sec. 19.10), we shall be dealing in peak-to-peak values of voltage and current. These are respectively double the single-peak values $E_{\max}$ and $I_{\max}$, and so power is equal to one eighth the product of the peak-to-peak values.

The r.m.s. value is *not* the same as the average voltage or current (which is actually 0.637 times the peak value, if averaged over a half cycle, or zero over a whole cycle). If you have followed the argument carefully you will see that this is because we have been comparing the a.c. and d.c. on a power basis, and power is proportional to voltage or current *squared.*

Remember that the figures given above apply only to the sine waveform. With a square wave (Fig. 6.2*b*) for example, it is obvious that peak, r.m.s., and average (or mean) values are all the same.

There is one other recognised 'value'—the instantaneous value, changing all the time in an a.c. system. It is the quantity graphed in Fig. 5.2. It can be found for any stage of a cycle by dividing the cycle into 360 degrees of angle as in Sec. 5.7 and then looking up the sine of that angle in a trigonometrical table and multiplying the peak value by it.

As regards symbols, plain capital letters such as E, V and I are generally understood to mean r.m.s. values unless the contrary is obvious. Instantaneous values, when it is necessary to distinguish them, are denoted by small letters, such as i; peak values by $I_{\max}$; and average (or mean) values by I_{ave}.

Using r.m.s. values for voltage and current we can forget the rapid variations in instantaneous voltage, and, *so long as our circuits are purely resistive,* carry out all a.c. calculations according to the rules discussed in Chapter 2.

5.5 A.C. Meters

If alternating current is passed through a moving-coil meter (Sec. 4.2) the coil is pushed first one way and then the other, because

5.6 Phase

Looking again at the current graph in Fig. 5.2 we see that it not only has the same shape as the e.m.f. that causes it but it is also exactly in step with it. Consideration of Ohm's law proves that this must be so in any purely resistive circuit. The technical word for being in step is being *in phase.* The idea of phase is very important, so we had better make sure we understand it.

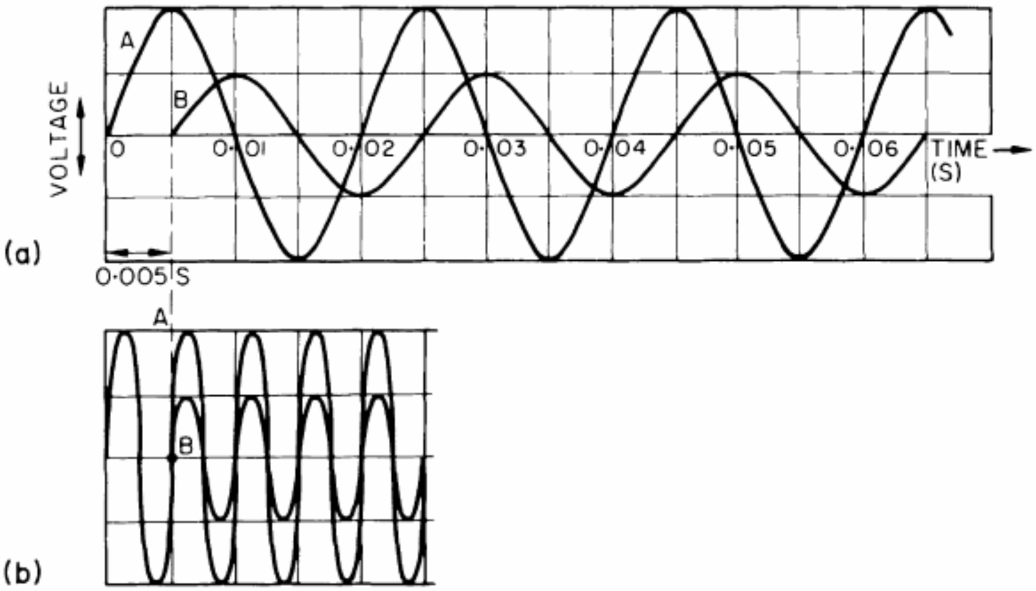


Fig. 5.3 The time lag between the start of A and the start of B is the same in *(a)* and *(b)*, but in *(a)* the two voltages are quarter of a cycle out of phase and in *(b)* are in phase. Time is therefore not the basic measure of phase

Suppose we have two alternating generators, A and B. The exact voltages they give will not matter, but to make it easier to distinguish their graphs we shall suppose A gives double the voltage of B. If we start drawing the voltage graph of A just as it begins a cycle the result will be something like waveform A in Fig. 5.3*a*. Next, consider generator B, which has the same frequency, but is out of step with A. It starts each of its cycles, say, quarter of a cycle later than A, as shown by waveform B. This fact is stated by saying that voltage B lags voltage A by a phase difference of quarter of a cycle or 90°. It may alternatively be expressed by saying that the two wave-forms are in *quadrature.* Seeing that it is a *time* graph, it might seem more natural to say that B lagged A by a phase difference of 0.005 s. The incorrectness of doing so is shown in Fig. 5.3*b*, where the frequency of A and B is four times as great. The time lag is the same as before, but the two voltages are now *in phase.* So although phase is usually very closely related to time, it is not advisable to think of it as time. The proper basis of reckoning phase is in fractions of a cycle.

5.7 Phasor Diagrams

The main advantage of a time graph of alternating voltage, current, etc., is that it shows the waveform. But on those very many occasions when the form of the wave is not in question (because we have agreed to stick to sine waves) it is a great waste of effort to draw a number of beautifully exact waves merely in order to show the relative phases. Having examined Figs. 1.2, 5.2 and 5.3, we ought by now to be able to take the sine waveform for granted, and be prepared to accept a simpler method of indicating phases.

In Fig. 5.4, imagine the line OP to be pivoted at O and to be capable of rotating steadily in the direction shown (anticlockwise), which by mathematical convention is the positive direction of rotation. Then the height of the end P above a horizontal line through O varies in exactly the same way as a sine wave. Assuming OP starts, as shown in Fig. 5.4, at '3 o'clock', P is actually on the horizontal line, so its height above it is, of course, nil. That represents the starting point of a sine-wave cycle. As OP rotates, its height above the line increases at first rapidly and then more slowly, until, after a quarter of a revolution, it is at right angles to its original position. That represents the first quarter of a sine-wave cycle. During the third and fourth quarter of the revolution, the height of P is negative, corresponding to the same quarters of a cycle.

Fig. 5.4 Starting position of an alternative method of representing sinusoidal variation

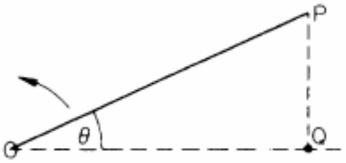


Fig. 5.5 The Fig. 5.4 phasor after having turned through an angle, θ

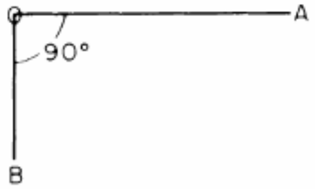


Fig. 5.6 Phasor diagram corresponding to Fig. 5.3*a*

Readers with any knowledge of trigonometry will know that the ratio of the height of P to its distance from O (i.e., PQ/PO in Fig. 5.5) is called the sine of the angle that OP has turned through; hence the name given to the waveform we have been considering and which is none other than a time graph of the sine of the angle θ (abbreviated 'sin θ') when θ (pronounced 'theta') is increasing at a steady rate.

We now have a much more easily drawn diagram for representing sine waves. The length of a line such as OP represents the peak value of the voltage, etc., and the angle it makes with the '3 o'clock' position represents its phase. The line itself is called a *phasor*. The others as the cycle progresses.

It is an essential part of the rotating phasor concept that the length of the phasor represents the peak value. But without forgetting this there is really no reason why we should not make use of the fact that the ratio between peak and r.m.s. values is constant ($\sqrt{2}$) to regard the phasors alternatively as representing r.m.s. values. It is merely a matter of choice of scale. This is convenient and is often done, especially when one has become used to interpreting stationary diagrams.

5.8 Adding Alternating Voltages

A great advantage of the phasor diagram is the way it simplifies adding and subtracting alternating voltages or currents that are not in phase. To add voltages A and B in Fig. 5.3*a*, for example, one would have to add the heights of the A and B waves at close intervals of time and plot them as a third wave. This would be a very laborious way of finding how the peak value and phase of the combined voltage (or *resultant,* as it is called) compared with those of its component voltages. In a phasor diagram it is quickly done by 'adding' phasor B to the rotating end of phasor A, as in Fig. 5.7, and drawing a straight line from 0 to the 'loose' end of B. This line is the phasor of the combined voltage, A + B. Its length represents its peak value (to the same scale as A and B) and its angle represents its relative phase. Alternatively of course A can be added to B. Note that the result of adding two voltages that are not in phase is less than the result of adding them by ordinary arithmetic, and its phase is somewhere between those of the components. Since it is representable by a phasor, it is sinusoidal, like its components.

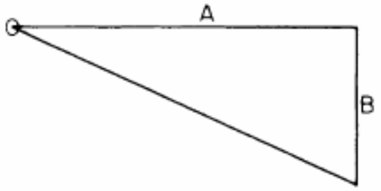


Fig. 5.7 How to find the phasor representing the resultant of A and B (Fig. 5.3*a*)

5.9 Direction Signs

Have you ever been directed to a destination by someone who supposed that it was being approached from the opposite direction? It is a very perplexing and frustrating experience. Yet quite a common one for students of a.c., because much of the literature on the subject is ambiguous in its direction signs.

All our talk about voltages and currents will do us no practical good unless we can relate them to circuits. Already we have been drawing graphs and phasor diagrams of voltages A and B without any reference to where they could be found. So let us suppose that voltage A (call it V_A) occurs between terminals c and d in a circuit, and V_B between terminals e and f in the same or another circuit. As these circuits might be very complicated, in Fig. 5.8 only the terminals are shown. A very common way of indicating the voltages between them is as shown there, with two-way arrows. Now let us try to relate this to the waveform graph, Fig. 5.3*a*, which tells us that the positive peak of V_B occurs quarter of a cycle later than the positive peak of V_A. The phasor diagram, Fig. 5.6, tells us the same thing in a simpler way. Suppose that at a certain instant d is at positive peak relative to c. Then we know that a quarter of a cycle later either e or f will be at peak positive relative to the other. Which one? The answer we have is like that of the man who was asked 'Will Oxford or Cambridge win the boat race?' and replied 'Yes'. Quite true, but quite useless. So are two-way arrows as voltage signposts.

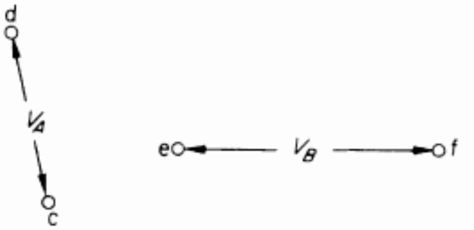


Fig. 5.8 If V_B is known to lag quarter of a cycle behind V_A, which of its terminals first becomes positive after d?

One solution to the difficulty is to insist on arrows pointing one way only, like any sensible signpost. The same applies of course to currents. When we were dealing with d.c. circuits, such as Fig. 2.14, one point that had to be made clear was that there needed to be agreement on whether the arrow heads marked what we knew were (say) the positive ends, or merely the ends that were to be taken as positive, so that if any voltage, at first unknown, turned out to be negative we would know that the arrow head would mark what was actually the negative end. With a.c. the voltages and currents are continually reversing, so no terminal or direction is any more positive than negative; but one-way arrows or even + and – signs can be used for a.c. to show the polarity that is arbitrarily taken as positive at a given instant. Half a cycle later everything will be reversed, but things that were opposite in sign will still be so and like signs will still be like. Provided that all phasors are correctly marked to correspond (hence the arrow-head custom) this method can be made to work. *Phasor Diagrams* mentions eight disadvantages it has, however, and advises a system that is free from all of them. A brief outline follows.

5.10 Subscript Notation

If the terminals are labelled, as in Fig. 5.8, there is no need to mark the voltages or their directions on the circuit diagram at all: V_A is we choose to regard as the reference point, or zero potential, with respect to which all other voltages are reckoned, is the appropriate

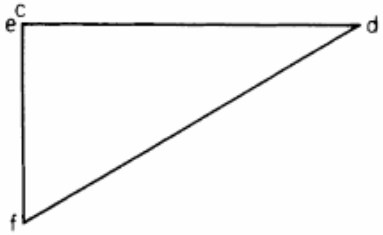


Fig. 5.10 If one of the circuit connections was reversed, this would apply instead of Fig. 5.9

centre, and we can mark it with a circle or stick a pin through it to serve as an axis.

5.11 Current Phasors

Attempts to extend subscript notation to currents usually run into trouble, because people write I_{cd} to mean the current flowing from c to d. But whereas there can be only one voltage at a time between two points, there can be as many currents as there are parallel paths. This difficulty can be avoided by labelling the *meshes* or loops of the circuit. There are excellent theoretical reasons for doing this, but it also works in practice. To distinguish currents from voltages, capital letters can be used for the meshes. All we need is to agree on which order of letters to use in order to denote the positive direction of current. Either would do, but to agree with *Phasor Diagrams* let us say that I_{AB} (which we can conveniently write as AB) in Fig. 5.11*a* means positive current moving up through the generator and down through the resistor. Note that although this simplest of circuits appears to have only one mesh we must mark the external one (A) too. Note also that I_B, or B, means the current circulating clockwise around mesh B, and AB means the net current crossed in moving from A to B.

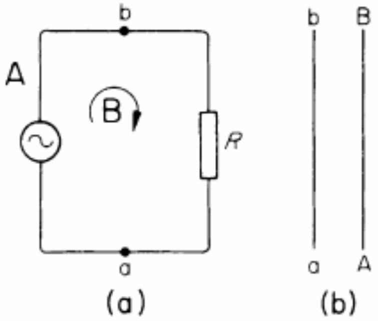


Fig. 5.11 Voltage and current phasor diagrams relating to the very simple circuit shown

For such a simple circuit the phasor diagram is very simple. There are only two terminals, so only one voltage, *ab* or *ba*, between them and therefore one voltage phasor (Fig. 5.11*b*). When *ab* is peak

CHAPTER 6

Capacitance in A.C. Circuits

6.1 Current Flow in a Capacitive Circuit

The next type of basic circuit to consider is the one shown in Fig. 6.1. If the generator gave an unvarying e.m.f. no current could flow, because there is a complete break in the circuit. The most that could happen would be a momentary current when switching on, as shown in Fig. 3.6.

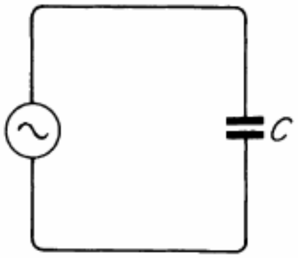
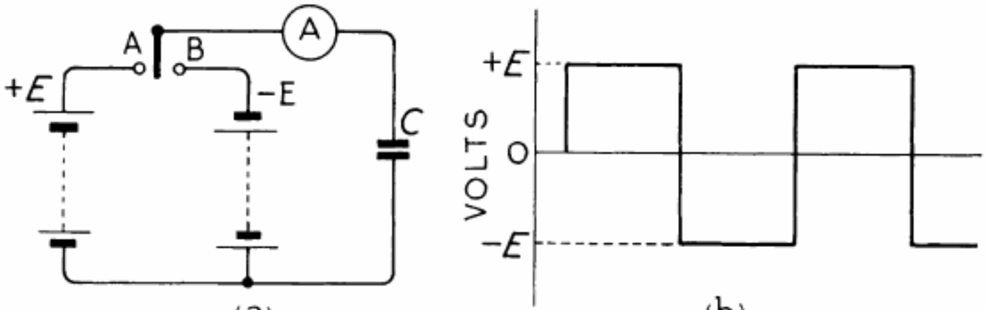


Fig. 6.1 Circuit consisting of an a.c. generator with a purely capacitive load, for comparison with Fig. 5.1

Some idea of what is likely to happen when an alternating e.m.f. is applied can be obtained by an experiment similar to Fig. 3.5 but with an inverted battery in the B path, as shown in Fig. 6.2*a*. Moving the switch alternately to A and B at equal time intervals will then provide an alternating voltage of square waveform (Fig. 6.2*b*).



(a) (b)

Fig. 6.2 If the generator in Fig 6.1 consisted of the periodically switched batteries as shown at (*a*), the waveform would be as at (*b*), and the current (assuming a certain amount of resistance) could be calculated as discussed in connection with Fig. 3.6

Whenever the switch is moved to A there is a momentary charging current in one direction, and moving it to B causes a similar current in the reverse direction.

We know that for a given voltage the quantity of electricity transferred at each movement of the switch is proportional to the capacitance (Sec. 3.2). And it is obvious that the more rapidly the switch is moved to and fro (that is to say, the higher the frequency of the alternating e.m.f.) the more often this quantity of electricity will surge to and fro in the circuit (that is to say, the greater the quantity of electricity that will move in the circuit per second, or, in other words, the greater the current).

So although the circuit has no conductance an alternating e.m.f. causes an alternating current. It does not flow conductively *through* the capacitor, whose dielectric is normally a very good insulator, but can be visualised as flowing in and out of it. Alternatively, Sec. 3.8 suggested how the current can be regarded as existing even between the capacitor plates.

The fact that an alternating e.m.f. can make a current flow in a circuit blocked to d.c. by a capacitance in series is easily demonstrated by bridging the open contacts of an electric light switch by a capacitor (of suitable a.c. rating!); say 2 μF for a 40 W lamp (Fig. 6.3). The lamp will light and stay alight as long as the capacitor is connected. But it will not be as bright as usual. The less the capacitance, the dimmer the light. The amount of current due to a given voltage thus depends on two things: the capacitance and the frequency.

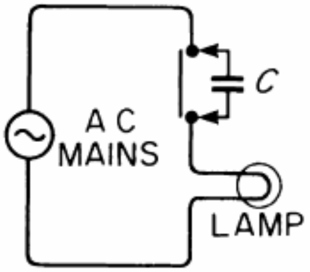


Fig. 6.3 Demonstration of current without conduction

6.2 Capacitive Current Waveform

Let us now consider the action of Fig. 6.1 in greater detail, by drawing the graph of instantaneous e.m.f. (Fig. 6.4) exactly as for the resistive circuit, in which we assumed a sine waveform. But unlike the resistive case there is no Ohm's law to guide us in plotting the current waveform. We have, however, a rather similar relationship (Sec. 3.2):

$$V = Q/C$$

where V is the p.d. across C, and Q is the charge in C. This is true at every instant, so we can rewrite it $v = q/C$, to show that we mean instantaneous values. In Fig. 6.1, v is always equal in magnitude to e, the instantaneous e.m.f. So we can say with confidence that at the moments when e is zero the capacitor is completely uncharged, and at all other moments the charge is exactly proportional to e. If we knew the right scale, the voltage wave in Fig. 6.4 would do also as a charge curve. But we are not so much interested in the charge as in the current; that is to say, the rate at which charging takes place. At points marked *a,* although the voltage and charge are zero they are growing faster than at any other stage in the cycle. So we may expect the current to be greater than at any other times. At points marked *b,* the charge is decreasing as fast as it was growing at *a*, so the current is the same in magnitude but opposite in direction and sign. At points marked *c,* the charge reaches its maximum, but just for an instant it is neither growing nor waning; its rate of increase or decrease is zero, so at that instant the current must be zero. At intermediate points the relative strength of current can be estimated from the steepness of the voltage (and charge) curve. Joining up all the points gives a current curve shaped like the dashed line.

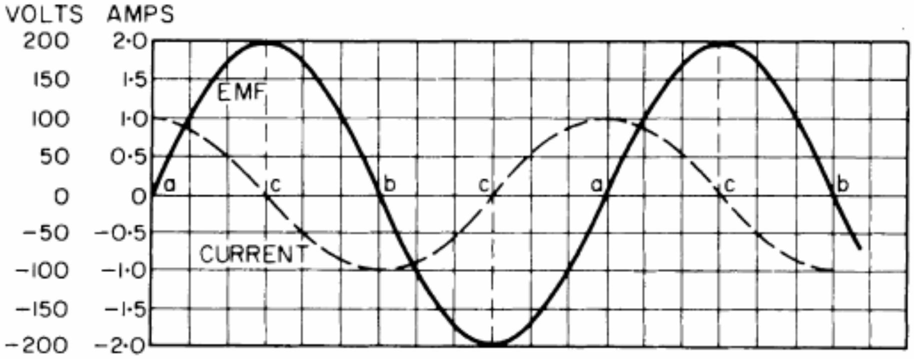


Fig. 6.4 E.m.f. and current diagram for Fig. 6.1, drawn to correspond with Fig. 5.2

If this job is done carefully the shape of the current curve is also sinusoidal, as in the resistive circuit, but out of phase, being a quarter of a cycle (or 90°) ahead of the voltage.

On seeing this, students are sometimes puzzled and ask how the current *can* be ahead of the voltage. 'How does the current know what the voltage is going to be, quarter of a cycle later?' This difficulty arises only when it supposed, quite wrongly, that the current maximum at *a* is caused by the voltage maximum a quarter of a cycle later, at *c*; whereas it is actually caused by the voltage at *a* increasing at its maximum rate.

6.3 The 'Ohm's Law' for Capacitance

What we want to know is how to predict the actual magnitude of the current, given the necessary circuit data. Since, as we know (Sec. 3.2), when there is 1 V between the terminals of a 1 F capacitor its charge must be 1 C, we can see that if the voltage were increasing at the rate of 1 V/s the charge would be increasing at 1 C/s. But we also know that 1 C/s is a current of 1 A; so from our basic relationship, $Q = VC$, we can derive

$$\text{Amps} = \text{Volts-per-second} \times \text{Farads}.$$

and then take another look at one of the forms of expressing Ohm's law:

$$\frac{E}{I} = R$$

we see they are the same except that $1/2\pi f C$ takes the place of R in the role of limiting the current in the circuit. So, although $1/2\pi f C$ is quite a different thing physically from resistance, for purposes of calculation it is of the same kind, and it is convenient to reckon it in ohms. To distinguish it from resistance it is named *reactance.* To avoid having to repeat the rather cumbersome expression $1/2\pi f C$, the special symbol X has been allotted. The 'Ohm's law' for the Fig. 6.1 circuit can therefore be written simply as

$$\frac{E}{I} = X \quad \text{or} \quad I = \frac{E}{X} \quad \text{or} \quad E = IX$$

For example, the reactance of a 2 μF capacitor to a 50 Hz supply is $1/(2\pi \times 50 \times 2 \times 10^{-6}) = 1\,592\,\Omega$, and if the voltage across it were 240 V the current would be $240/1592 = 0.151$A.

The factor group $2\pi f$ occurs so often in connection with a.c. circuits that it is commonly denoted by the Greek small omega, ω. So an alternative way of writing the reactance is $1/\omega C$. But in this elementary book the formulae will be used mainly for numerical examples, so it may be clearer if the separate factors are shown.

6.4 Capacitances in Parallel and in Series

The argument that led us to conclude that the capacitance between parallel plates is proportional to the area across which they face one another (Sec. 3.3) leads also to the conclusion that capacitances in parallel add up just like resistances in series.

We can arrive at the same conclusion by simple algebraic reasoning based on the behaviour of capacitors to alternating voltage. Fig. 6.6 represents an a.c. generator connected to two capacitors in parallel. The separate currents in them are respectively $E \times 2\pi f C_1$ and $E \times 2\pi f C_2$. As these currents are in phase with one another (because both lead E by the same angle), the total current is $E \times 2\pi f (C_1 + C_2)$, which is equal to the current that would be taken

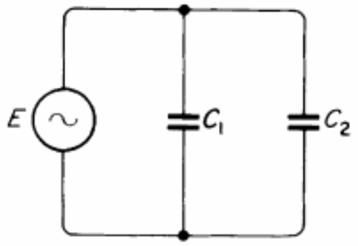


Fig. 6.6 In this circuit the capacitance of C_1 and C_2 together is equal to $C_1 + C_2$

6.5 Power in a Capacitive Circuit

Fig. 6.4 was drawn to match Fig. 5.2 as regards peak voltage and current. The only difference is the current's 90° phase lead. Let us now complete the picture by calculating the wattage at a sufficient number of points to draw the power curve (Fig. 6.8). Owing to the phase difference between *E* and *I*, there are periods in the cycle when they are of opposite sign, giving a negative power. In fact, the power is alternately positive and negative, just like *E* and *I* except for having twice the frequency. The net power, taken over a whole cycle, is therefore zero. Since we are applying the sign '+' to positive current going through the generator towards positive potential, positive power means power going out of the generator, and negative power is power returned to the generator by the load (end of Sec. 2.17).

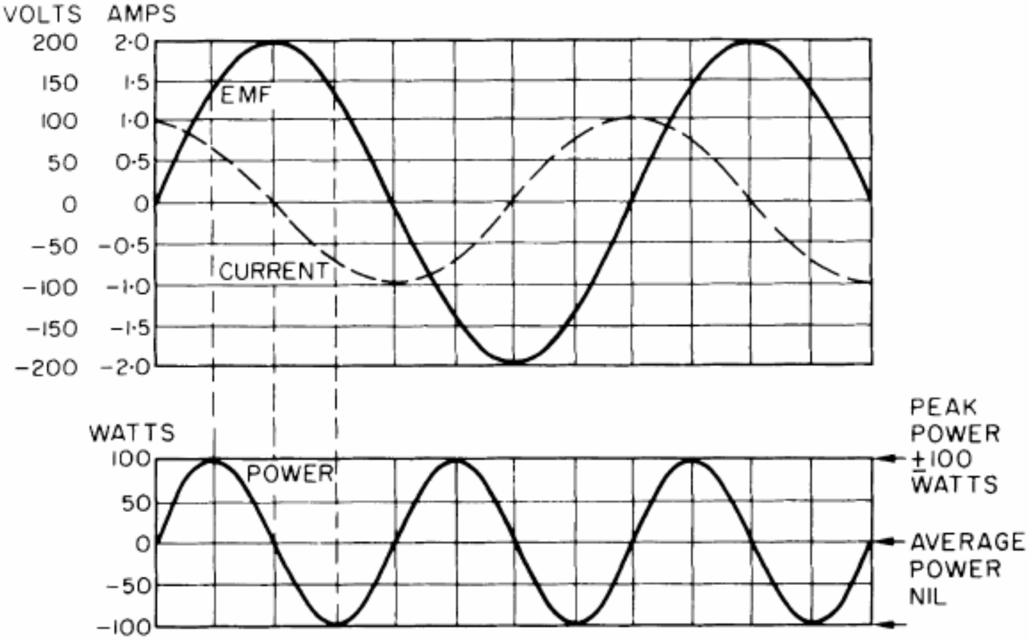


Fig. 6.8 The current and voltage graphs for a capacitance load (Fig. 6.4) are here repeated, with a power graph derived from it plotted below

The power curve in Fig. 6.8 represents the fact that although a pure capacitance draws current from an alternating generator it does not permanently draw any power—it only borrows some twice per cycle, repaying it in full with a promptitude that might be commended to human borrowers.

A demonstration of this fact is given by the capacitor in series with the lamp (Fig. 6.3). The lamp soon warms up, showing that electrical power is being dissipated in it, but although the capacitor is carrying the same current it remains quite cold.

6.6 Phasor Diagram for Capacitive Circuit

In dealing with such a simple circuit as Fig. 6.1 we are not likely to become uncertain about the directions of e.m.f. and current represented in graphs such as Fig. 6.8, especially as we have just stated them fully in words in order to check the directions of power flow. But in order to get into practice for more complicated circuits let us draw the phasors for Fig. 6.1, as we did for the simple resistive circuit in Fig. 5.11.

Fig. 6.9*a* shows the circuit suitably labelled. We can always choose the angle at which to draw the first phasor, so to make it easy to compare with Fig. 5.11 let us draw the voltage phasor in the same way, indicating that terminal b is at peak positive potential with respect to a. At that moment, therefore, the voltage across *C* has just stopped increasing and is about to decrease; so *C* is charged to the maximum extent, positive on the upper plate. The current, which has been flowing clockwise to deliver that charge, is now at zero and about to discharge anticlockwise. The current phasor must therefore be drawn horizontal, to signify zero current. And as anticlockwise current, which our convention denotes by *BA,* is the one about to become positive, it must be drawn left to right as shown (Fig. 6.9*b*). As *ab* is the generator voltage applied to *C,* one would usually refer to the clockwise current *AB,* which the diagram shows to be a quarter of a cycle ahead of *ab.* This confirms the rule for sinusoidal a.c., well known to electrical engineers, that the current into a pure capacitance leads the applied voltage by 90°.

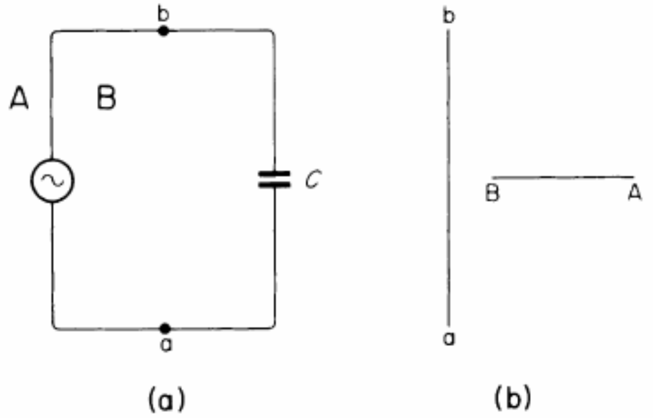


Fig. 6.9 Simplest capacitive circuit and corresponding phasor diagrams

6.7 Capacitance and Resistance in Series

Fig. 6.10 shows the next type of circuit to be considered: *C* and *R* in series. We know how to calculate the relationship (magnitude and phase) between voltage and current for each of these separately. When they are in series the same current will flow through both, and

You may ask how we know how long to draw these phasors. The current one has to be drawn first, and at that stage we do not know how much current there is, even if we know ab, f, C and R. Whatever AB may turn out to be, however, any length of phasor will be right if a suitable scale is chosen. As for ac and cb, those voltages are proportional to R and X ($= 1/2\pi fC$), so we can draw their phasor lengths in the ratio $R:X$. Then if we know the generator e.m.f., ab represents it, so the voltage scale is found and hence the values of ac and cb. Lastly, AB can be calculated by Ohm's law as ac/R.

Most of the time, however, phasor diagrams are drawn just to show the principle without bothering about the exact values they represent. At low frequencies at least, it is likely that C would impede current more than R and so take the major share of the available e.m.f., so cb has been drawn longer than ac.

6.8 Impedance

A new name is needed to refer to the current-limiting properties of this circuit as a whole. *Resistance* is appropriate to R and *reactance* to C; a combination of them is called *impedance* (symbol: Z). Resistance and reactance themselves are special cases of impedance. So we have still another relationship in the same form as Ohm's law; one that covers the other two, namely:

$$I = \frac{E}{Z}$$

Although Z combines R and X, it is not true to say $Z = R + X$. Suppose that in Fig. 6.10 C is the 2 μF capacitor which we have already calculated has a reactance of 1 592 Ω at 50 Hz, and that R is a round 1000 Ω. AB, shall we say, is 0.1 A. Then ac must be 100 V and cb 159 V. If the phasor diagram was drawn accurately to scale it would show ab to be 188 V, so Z must be 1880 Ω, which is much less than 1592 + 1000.

As the three voltage phasors refer to the same current, to a suitable scale they also represent the corresponding three impedances: resistive, reactive and total. Also, as they make a right-angled triangle, the celebrated Theorem of Pythagoras tells us that

$$Z^2 = R^2 + X^2$$

$$\text{or} \quad Z = \sqrt{(R^2 + X^2)}$$

In trigonometry the ratio cb/ac, which we have just seen is equal to X/R, is known as the tangent of the angle φ, abbreviated to tan φ. As there are tables from which the angle corresponding to any tangent can be seen and some pocket-calculators include them, phase angles can be found from

$$\tan \varphi = \frac{X}{R}$$

So we have the alternative of computing Z by arithmetic instead of drawing a phasor diagram or a waveform diagram accurately to scale.

The phasor triangle also tells us, of course, that

$$ab = \sqrt{(ac^2 + cb^2)}$$

(Note that here *'ac'* is really one symbol, all squared; not $a \times c^2$.) We will no doubt have already realised that 100 V and 159 V, the voltages across R and C, do not add up in the ordinary way to equal 188 V, the voltage across both. This would be very puzzling—in fact, quite impossible, and a serious breach of Kirchhoff's voltage law—if these figures were instantaneous values. But of course they are r.m.s. values, and the explanation of the apparent discrepancy is that the maximum instantaneous values do not occur all at the same time. This is most clearly shown by a waveform graph, tedious though it is to draw as a means of calculating the total voltage, etc., compared with a phasor diagram or algebra.

But this type of phasor diagram does very neatly illustrate Kirchhoff's voltage law, because the voltage phasors for any whole circuit always necessarily form a closed geometrical figure, such as our triangle.

6.9 Power in Mixed Circuits

By now we should hardly need to draw waveform diagrams such as Fig. 6.8 in order to find how much power the generator delivers to the Fig. 6.10 circuit. Fig. 6.8, together with Fig. 6.9, has shown that when phasors are at right angles, as they are when they refer to a perfect capacitor, energy flows to and fro between generator and load, but none is permanently delivered. That is because a capacitor can store electrical energy and returns it in the same form, unlike a resistor, which dissipates it all as heat just as fast as it is received. In Sec. 5.4 we found that the power P delivered to and dissipated by resistance R is E^2/R, where E is the r.m.s. voltage across it. As $R = E/I$, I being the r.m.s. current through R, by substituting for R in the first equation we get $P = EI$, just as with d.c. (Sec. 2.17). The quantities corresponding in Fig. 6.10 to E and I are AB and ac, if those are being used to represent r.m.s. values. If however they are used more strictly to represent peak values, and since peak values are $\sqrt{2} \times$ r.m.s. values, P would be equal to $AB.\ ac/2$.

In a circuit such as this, including both R and C in series, the part of the total voltage that should be multiplied by the current to find the power dissipated is represented by the phasor that is parallel to the current phasor: ac in Fig. 6.10*b*. In terms of the total voltage, this voltage is *ab*. *ac*/*ab*, or *ER*/*Z*. So

$$P = \frac{EIR}{Z}$$

E and I being the total voltage and the current.

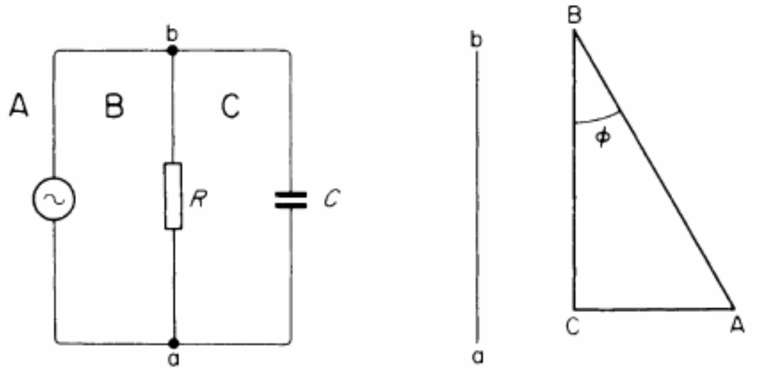
In trigonometry the ratio ac/ab is called the cosine of the angle φ, abbreviated to cos φ. So another form of the same equation is

$$P = EI \cos \varphi$$

φ being the phase angle between E and I. Cos φ is known to electrical engineers as the *power factor.*

6.10 Capacitance and Resistance in Parallel

In solving this type of circuit, Fig. 6.11, we revert to the practice of starting with the voltage, because that is common to both. What we do, in fact, is exactly the same as for the series circuit except that currents and voltages change places. We now have two currents, *CB* and *AC*, one in phase with the e.m.f. *ab* and the other 90° ahead of it. By drawing the phasors accordingly, Fig. 6.11*b*, we find the total current, *AB*, and the phase angle φ by which it leads *ab*.



(a) (b)

Fig. 6.11 A parallel circuit such as this is similar to a series circuit such as Fig. 6.10 with voltages and currents interchanged

By the same methods as we used in Sec. 6.8 we find that

$$AB = \sqrt{(AC^2 + CB^2)}$$

This is in fact the same equation, apart from the capital letters for currents instead of small ones for voltages. As with them, *AC* is regarded as one symbol, all squared.

Calculating the impedance is slightly more complicated, however, because the currents are *inversely* proportional to the impedances

CHAPTER 7

Inductance in A.C. Circuits

7.1 Current Flow in an Inductive Circuit

Although the phenomenon of magnetism, which gives rise to inductance, differs in many ways from electrostatics, which gives rise to capacitance, it is helpful to draw a very close parallel or analogy between them. Inductance and capacitance are, in fact, like opposite partners; and this chapter will in many ways be a repetition of the last one, but with a few basic things reversed.

We have already noticed some striking similarities as well as differences between capacitance and inductance (Sec. 4.10). One thing that could be seen by comparing them in Figs. 3.6 and 4.10 was an exchange of roles between voltage and current.

The simple experiment of Fig. 6.2 showed that, in a circuit consisting of a square-wave alternating-voltage generator in series with a capacitor, the current increased with frequency, beginning with zero current at zero frequency. This was confirmed by examining in more detail a circuit (Fig. 6.1) with a sine-wave generator, in which the current was found to be exactly proportional to frequency.

Comparing Fig. 4.10 with Fig. 3.6, we can expect the opposite to apply to inductance. At zero frequency it has no restrictive influence on the current, which is limited only by the resistance of the circuit. But when a high-frequency alternating e.m.f. is applied the current has very little time to grow (at a rate depending on the inductance) before the second half-cycle is giving it the 'about turn'. The gradualness with which the current rises, you will remember, is due to the magnetic field created by the current, which generates an e.m.f. opposing the e.m.f. that is driving the current. It is rather like the gradualness with which a heavy truck gains speed when you push it. If you shake it rapidly to and fro it will hardly move at all.

For finding exactly how much current a given alternating voltage will drive through a given inductance (Fig. 7.1) we have the basic relationship (Sec. 4.4):

Volts = Amps-per-second × Henries (Compare Sec. 6.3.)

Adapting the same basic method we begin with a sinusoidal current and find what voltage is required. The dashed curve in Fig. 7.2 represents such a current during rather more than one cycle. At the start of a cycle (*a*), the current is zero, but is increasing faster than at any other phase. So the amps-per-second is at its maximum, and so therefore must be the voltage. After a half a cycle (*b*) the current is again zero and changing at the same rate, but this time it is *decreasing,* so the voltage is a maximum *negative.* Halfway between (*c*) the current is at its maximum, but for an instant it is neither increasing nor decreasing, so the voltage must be zero. And so on. Completing the voltage curve in the same way as we did the current curve in Fig. 6.4, we find that it also is sinusoidal. This is very fortunate, for it allows us to say that if the sinusoidal voltage represented by this curve were applied to an inductance, the current curve would represent the resulting current, which would be sinusoidal.

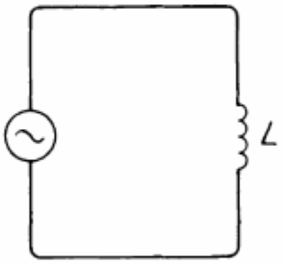


Fig. 7.1 Circuit consisting of an a.c. generator with a purely inductive load

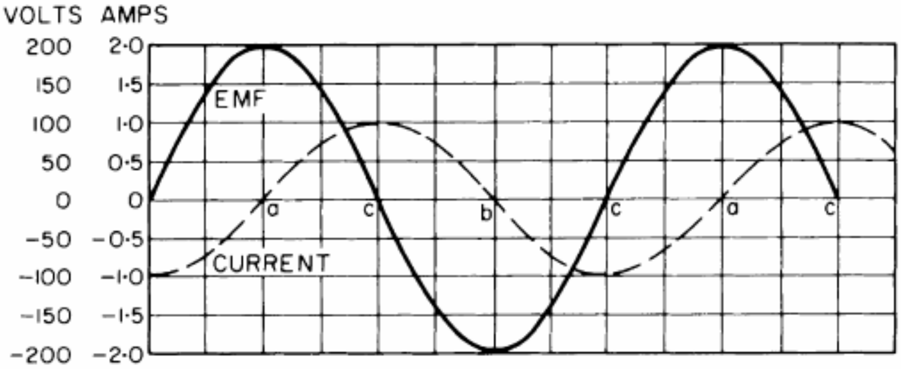


Fig. 7.2 E.m.f. and current graphs for Fig. 7.1, drawn to correspond with Figs. 5.2 and 6.4

7.2 The 'Ohm's Law' for Inductance

We have already found (Sec. 6.3) that when an alternating e.m.f. E, of frequency f, is sinusoidal, its r.m.s. volts-per-second is $2\pi fE$. The same method applies equally to current, so the r.m.s. amps-per-second is $2\pi fI$. Fitting this fact into our basic principle we get

$$E = 2\pi fIL$$

where L is the inductance in henries.

What we set out to find was the current due to a given e.m.f., so the appopriate form of the above equation is

$$I = \frac{E}{2\pi fL}$$

Again, this is the same form as Ohm's law, $2\pi fL$ taking the place of R. This $2\pi fL$ can therefore also be reckoned in ohms, and, like $1/2\pi fC$, it is called reactance and denoted by X. Whenever it is necessary to distinguish between inductive reactance and capacitive reactance they are denoted by X_L and X_C respectively. X is the general symbol for reactance.

A question that may come to mind at this point is: can X_L and X_C in the same circuit be added in the same simple way as resistances? The answer is so important that it is reserved for the next chapter.

Note in the meantime that X_L is proportional to f, whereas X_C is proportional to $1/f$; this expresses in exact terms what our early experiments had led us to expect about the opposite ways in which frequency affected the current due to a given voltage.

An example may help to clinch the matter. What is the reactance of a 2 H inductor at 50 Hz? $X = 2\pi fL = 2\pi \times 50 \times 2 = 628\ \Omega$. So if the voltage is 240, the current will be $240/628 = 0.382$ A.

7.3 Inductances in Series and in Parallel

The previous section has shown that the reactive impedance of an inductance is directly proportional to the inductance—oppositely to that of capacitance (which is proportional to $1/C$), but exactly like resistance. So inductances can be combined in the same way as resistances. The total effect of two in series, L_1 and L_2, is equal to that of a single inductance, $L_1 + L_2$. And inductances in parallel follow the reciprocal law: $1/L = 1/L_1 + 1/L_2$, or

$$L = \frac{L_1 L_2}{L_1 + L_2}$$

These principles can easily be verified by adding the reactances when they are in series, and the currents when the inductances are in parallel.

The above rules are subject to one important condition, however: that the inductors are so placed that their mutual inductance (Sec. 4.11) or magnetic coupling is negligible. If mutual inductance, M, does exist between two coils having separate self-inductances L_1 and L_1, the total self-inductance of the two in series is $L_1 + L_2 + 2M$ or $L_1 + L_2 - 2M$, according to whether the coils are placed so that their separate magnetic fields add to or subtract from one another. The reason M is doubled is that each coil is affected by the other, so in the combination the same effect occurs twice.

The corresponding elaboration of the formula for two inductances in parallel gives the result

$$L = \frac{(L_1 \pm M)(L_2 \pm M)}{L_1 + L_2 \pm 2M}$$

where '±' signifies '+ or −'.

Complications arise also when the fields of two capacitors interact, but with the usual forms of construction such interaction is seldom enough to take into account.

7.4 Power in an Inductive Circuit

Comparing Figs. 6.4 and 7.2 we see one difference. Whereas the capacitive current leads its terminal voltage by a quarter of a cycle (90°), the inductive current lags by that amount. If you want something to do you might care to calculate the instantaneous power at intervals throughout the cycle in Fig. 7.2 and draw a power curve, as in Fig. 6.8. But the same result can be achieved with less trouble simply by reversing the time scale in Fig. 6.8, making it read from right to left. So except for relative phases the conclusions regarding power in a pure inductance are exactly the same as for a capacitance: the net power taken over a whole cycle is zero, because during half the time the generator is expending power in building up a magnetic field, and during the other half the energy thus built up is returning power to the generator. So in the example we took, with 240 V passing 0.382 A through 2 H, 240×0.382 does not represent watts dissipated in the circuit (as it would with resistance) but half that number of volt-amps tossed to and fro between generator and inductor at the rate of 100 times a second.

If you try an experiment similar to Fig. 6.3, but using a coil of several henries in place of the capacitor, you may find that it does become perceptibly warm. This is because the coil is not purely inductive. Whereas the resistance of the plates and connections of a capacitor is usually negligible compared with its reactance, the resistance of the wire in a coil is generally very appreciable. Although the resistance and inductance of a coil are physically inseparable, it is allowable to consider them theoretically as separate items in series with one another.

7.5 Phasor Diagram for Inductive Circuit

For comparison with Fig. 6.9 we draw the simple circuit diagram, Fig. 7.3*a*, of a generator feeding a load consisting of pure (i.e., resistanceless) inductance. We start with the current phasor and by applying the basic principles we know, as in Sec. 6.6, we find that the correct relative angle of the voltage phasor is as shown (Fig. 7.3*b)*. With the same conventions as before, it expresses the well-known rule that the current through a pure inductance lags the applied sinusoidal voltage by 90°.

7.6 Inductance and Resistance in Series

R in Fig. 7.4*a* represents the resistance of the inductor whose inductance is represented by *L,* plus any other resistance there may be in series. This circuit can be tackled in exactly the same manner as Fig. 6.10. As the current is common to both *L* and *R* we again begin with its phasor. The voltage phasor ac is again parallel to AB, and in fact the only real difference is that the phasor representing the voltage *cb* applied to *L* must be drawn so that AB is 90° behind in phase, which is the same thing as saying cb must lead AB by that angle.

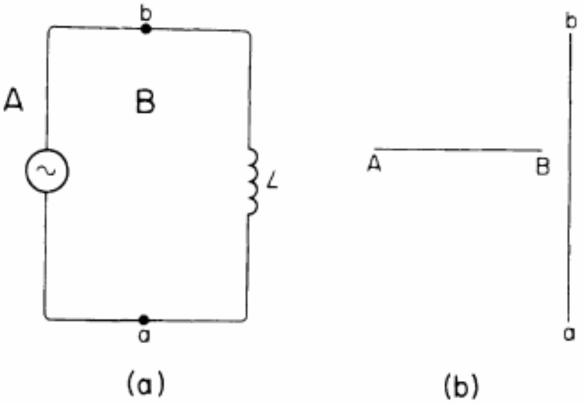


Fig. 7.3 The circuit of Fig. 7.1 repeated with its phasor diagrams

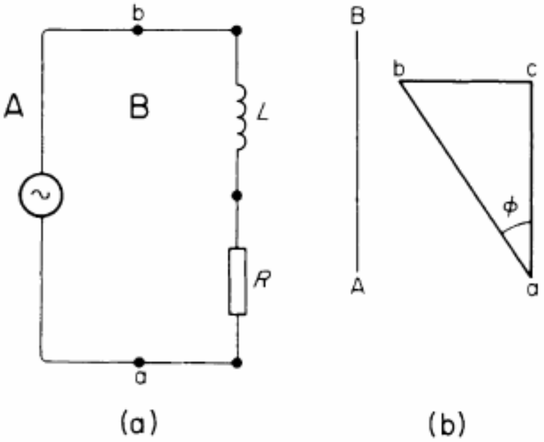


Fig. 7.4 *L* and *R* in series, and corresponding phasor diagrams

As ac is parallel to AB, φ is the angle by which *AB* lags the total applied voltage, *ab*. Again, as in Secs 6.8 and 6.9, and by similar reasoning,

$$Z = \sqrt{(R^2 + X^2)}$$

$$\tan \varphi = \frac{X}{R}$$

$$\text{and} \quad P = EI \cos \varphi$$

7.7 Inductance and Resistance in Parallel

Fig. 7.5 corresponds to Fig. 6.11 and the phasor diagram is derived in the same way. As the current AC through L lags the applied voltage ab by 90° instead of leading it as it would with a capacitor, AC must be drawn in the opposite direction, and φ is the angle by which the total current supplied by the generator lags its voltage.

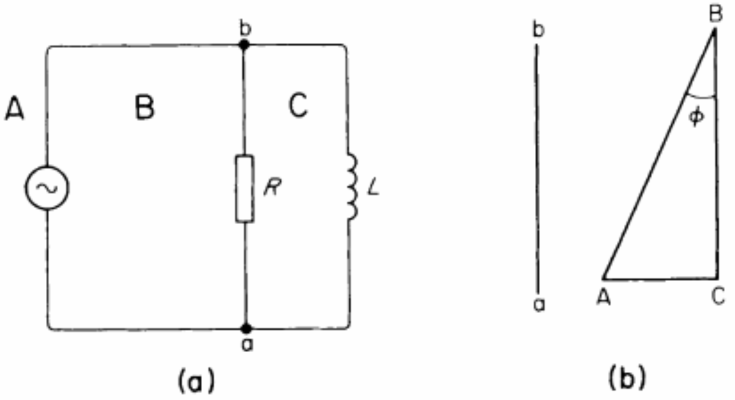


Fig. 7.5 Parallel *LR* circuit and phasor diagrams

If R is independent of L, then this circuit is somewhat unrealistic, as the resistance of the wire with which L is wound is not allowed for. If its resistance were included (in series with L) the problem could be solved by combining the methods for series and parallel impedances, as in Sec. 8.9. We shall see however, in Sec. 8.16, that Fig. 7.5*a* is an alternative way of taking account of the resistance of a reactive component.

The formulae for Z, Y, etc., found in Sec. 6.10, apply equally here, X standing for inductive instead of capacitive reactance.

The more advanced books extend the methods and scope of circuit calculation enormously; but before going on to them one should have thoroughly grasped the contents of these last two chapters. Some of the results can be summarised like this:

			Applied Voltage:
Impedance, Z ohms	Resistance, R ohms		In phase with current
	Reactance, X ohms	Capacitive, $X_C = 1/2\pi fC$	Lags current by 90°
		Inductive $X_L = 2\pi fL$	Leads current by 90°

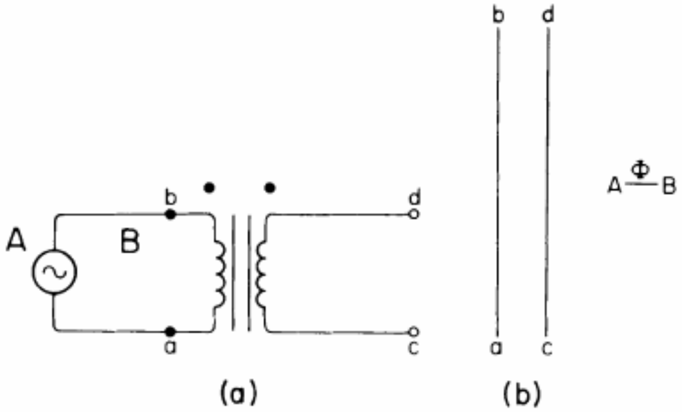
With R and X in series, $Z = \sqrt{(R^2 + X^2)}$ and $\tan\varphi = X/R$.

			Current:	
Admittance, Y siemens	Conductance, G siemens		In phase with applied voltage	
	Susceptance, B siemens	Capacitive, $B_C = 2\pi fC$	Leads	applied voltage by 90°
		Inductive, $B_L = 1/2\pi fL$	Lags	

With G and B in parallel, $Y = \sqrt{(G^2 + B^2)}$ and $\tan \varphi = B/G$.

7.8 Transformers

In Sec. 4.11 we had a brief introduction to the transformer, which is an appliance very widely used for electrical purposes, including radio and electronics. We are now equipped to study it in a little more detail, beginning with Fig. 7.6*a*. This represents a transformer connected to a generator giving the usual sinusoidal waveform. Let us assume that the primary and secondary coils have the same number of turns, and at this stage such complications as their resistance are conveniently assumed to be absent. An iron core is used to ensure that we can also assume, without serious error, that all the magnetic flux generated by the primary coil also links with the secondary, which at present is not connected to anything. The dots marking the top ends of the coils are a standard convention to exclude uncertainty about the relative direction of the windings. Without it one could not be sure about the polarity of the secondary voltage. If both coils are wound around the core in the same direction, then the dots mark both starting ends *or* both finishing ends.



(a) (b)

Fig. 7.6 Unloaded 1 : 1 transformer with 100% coupling

With the secondary coil open-circuited, as shown, the circuit is essentially the same as Fig. 7.3, so the phasor diagram is the same. We saw in Sec. 7.2 that the current driven by any applied voltage E

through an inductance *L* is inversely proportional to *L*. And in Sec. 4.6 we noted that the inductance of an ideal inductor of the type we are considering is directly proportional to the permeability, μ of the core, and to the square of *N*, the number of turns. The iron core of a transformer normally has a μ running into thousands, and *N* is usually considerable, so the inductance is likely to be very large and the current therefore quite small. That is why AB is drawn so short in Fig. 7.6*b*. *AB* is called the *magnetising current* because its purpose is to generate sufficient alternating magnetic flux in the core to induce an e.m.f. in the primary coil exactly equal to that applied. As the flux is due directly to the current (not to its rate of change), it is in phase with the current, so the same phasor will do for both. This is shown by marking it with the flux symbol, Φ (Greek capital phi) in Fig. 7.6*b*; and this also picks out *AB* as the magnetising current. Actually, as mentioned in Sec. 4.7, μ depends on the flux density in the core, varying throughout each cycle; so to induce a sinusoidal e.m.f.—necessary to match the applied one at every instant—the magnetising current has to be distorted, or non-sinusoidal, and therefore cannot strictly be represented by a phasor. This awkward complication is usually ignored, especially as the magnetising current is such a little one. With the ideal conditions assumed, it is clear that the average power required from the generator is nil, but that does not mean that there is no point in keeping the magnetising current as small as possible; the resistances of coil, generator and leads, which we are neglecting, give rise to a power loss proportional to the square of the current.

Because the secondary coil in the assumed transformer has the same number of turns as the primary and is linked by the same alternating flux, the secondary voltage *cd* must be the same as the primary, *ab*, so is represented by a phasor of the same length and angle. In fact, a single phasor would do, and this corresponds to the fact that a could be connected to c and b to d without affecting the electrical conditions. This is where the dots are helpful; if the secondary dot was at the c end it would indicate a reversal of winding and consequent polarity of secondary voltage, so paralleling the two coils as described would be disastrous!

7.9 Load Currents

The next step is to connect a load (for full explanation of this term, see Sec. 11.2) to the secondary terminals, and continuing our ideal simplicity we make it a resistor, R in Fig. 7.7*a*. This creates a new current mesh, C, and also raises a question about how this should be related to mesh B. Are we to suppose that the external mesh A penetrates between the windings, or should we think of them as too close together for that? According to our current notation *BC* is the

7.10 Transformer Losses

In real transformers the primary coil has some resistance as well as inductance. Moreover, the alternating flux in the core generates a certain amount of heat in it (Sec. 8.14), which has to come from somewhere, and in accordance with our wider definition of resistance (Sec. 2.18) can be considered as equivalent to some extra resistance in the primary. The primary resistance, augmented in this way, causes the magnetising current to lag the primary voltage by less than 90°.

Nor is the power from the supply transferred to the load without loss. The primary and secondary load currents have to pass through the resistances of the windings. These resistances are represented in Fig. 7.8 by R_1 and R_2. Moreover in real transformers some of the flux linking the primary coil fails to link with the secondary; similarly when secondary current is flowing it creates some flux not linked with the primary. These fluxes can be represented by small inductances L_1 and L_2, called *leakage inductances.* Note that the transformer terminals are still a-b and c-d; the internal junctions are of course not accessible as they are entirely imaginary. Fig. 7.8 is an example of an *equivalent circuit*—equivalent in this case, approximately, to a real transformer, so far as any actual electrical measurements can show. It happens to be an example of a very common type of equivalent circuit, with two input and two output terminals; and because one may have to guess, from external measurements, what circuit best represents its actual behaviour, it is often called a black box.

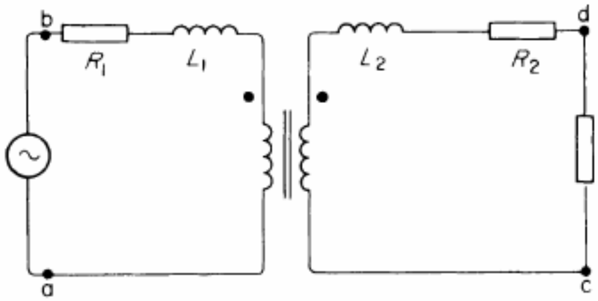
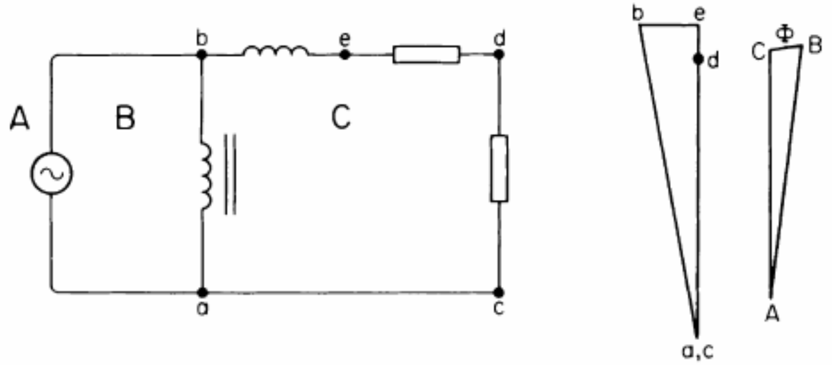


Fig. 7.8 Fig. 7.7*a* is here elaborated to include the resistances of the coils and their leakage inductances

If you feel confident about applying the phasor diagram rules, Fig. 7.8 would be a good exercise. Although more complicated than anything we have tackled so far, it is quite straightforward, if one labels the four extra junctions and starts with the *cd* phasor. However, as it may be too big a jump for some, let us deal with a simplified version.

We have already seen that if the transformer turns ratio is 1 : 1 the two coils can be reduced to one. And as the magnetising current is normally quite small compared with the load current its contribution to the voltage drop across R_1 and L_1 can reasonably be neglected, by transferring them to the secondary circuit, so that suitably increased values of R_2 and L_2 include them. Fig. 7.9*a* shows the simplified diagram. For easy comparison with the preceding ones, c is still shown, though it is now electrically the same as a.



(a) (b)

Fig. 7.9 Simpler phasor diagrams (*b*) than those applicable to Fig. 7.8 are obtained by transferring the primary resistance and leakage inductance to the secondary circuit and combining the fully coupled parts of the windings (*a*)

AC and cd are first drawn parallel to one another. If the load had not been purely resistive they would of course have had to be drawn with the appropriate phase angle between them. de is simply an extension of cd, but eb must lead it by 90° because *eb* is across an inductance. We now see *ab,* the input voltage, and can therefore draw CB at right angles to it to represent an ideal magnetising current, or at a somewhat smaller angle to be more realistic. As the diagram clearly shows, one effect of coil resistance and leakage inductance is that the output voltage falls as the current taken is increased. This effect is confirmed by practical experience. And if we calculate the output power, equal to *cd. AC,* and compare it with the input power, which in Sec. 6.9 we found to be *ab*. *AB* cos φ, we can find how much power is lost in the transformer. The ratio of output to input power, which is always less than 1, is called the *efficiency* of the transformer.

These considerations are most important in mains transformers; in audio-frequency transformers (which are less used than they once were) the emphasis is rather different, and in radio-frequency transformers it is almost completely different, as will be seen when they come up for attention.

7.11 Impedance Transformation

So far our transformer diagrams have been related to those having the comparatively little-used 1 : 1 ratio. There is no reason why other

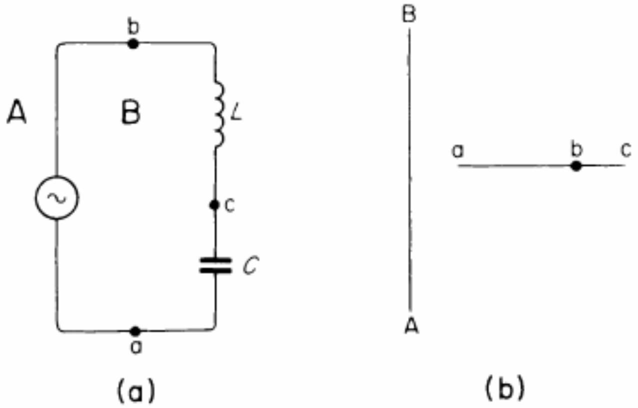
CHAPTER 8

The Tuned Circuit

8.1 Inductance and Capacitance in Series

In the previous chapters we have seen what happens in a circuit containing capacitance *or* inductance (with or without resistance), and the question naturally arises: what about circuits containing both? We know that two or more reactances of either the inductive or capacitive kind in series can be combined just like resistances, by adding; and either sort of reactance can be combined with resistance by the more complicated square-root process. The outcome of combining reactances of opposite kinds is so fundamental to radio and many other electronic applications that it needs a chapter to itself.

We shall begin by considering the simplest possible series circuit, Fig. 8.1 with capacitance and inductance in series. The method is exactly the same as for reactance and resistance in series; that is to say, since the current is common to both it is easiest to start from it and work backwards to find the e.m.f. needed to drive it. As in Fig. 7.4, we first draw a current phasor AB at any angle we like. The voltage applied to L (cb) must lead AB by 90°; ac on the other hand must lag AB by 90°. And so we get the complete phasor diagram. As a further check we may care to take another look at phasor and waveform diagrams for C alone (Figs. 6.9 and 6.4) and for L alone (7.3 and 7.2). The conclusion must be the same: that in Fig. 8.1 the voltages across C and L are in opposite phase, or 180° out of phase. So one of them must be *subtracted* from the other to give the



(a) (b)

Fig. 8.1 (*a*) circuit consisting of an a.c. generator in series with capacitance and inductance, and (*b*) corresponding phasor diagram necessary driving e.m.f. And since reactance is equal to applied voltage divided by current (the same in both cases) it follows that the total reactance of the circuit is also equal to one of the separate reactances *less* the other.

Which voltage or reactance must be subtracted from which to give the total? There is no particular reason for favouring either, but the agreed convention is to call X_L positive and X_C negative.

For example, suppose L in Fig. 8.1 is 2 H, C is 2μF, and *ab* is 240 V at 50 Hz. We have already calculated the reactances of 2 H and 2μF at 50 Hz (Secs. 7.2 and end of 6.3) and found them to be 628 Ω and 1592 Ω. So the total reactance must be $628 - 1\,592 = -964$ Ω, which at 50 Hz is the reactance of a 3.3 μF capacitor. The generator, then, would not notice any difference (at that particular frequency) if a 3.3 μF capacitor were substituted for the circuit consisting of 2 μF in series with 2 H. The magnitude and phase of the current would be the same in both cases; namely, $240/964 = 0.294$ A, leading by 90°.

To make sure, let us check it by calculating the voltage needed to cause this current. *cb* is equal to X_C times AB, $= 0.249 \times 1592 =$ 396 V. *ac* is X_L times AB, $= 0.249 \times 628 = 156$ V. The resultant of *ac* and *cb*, namely *ab*, must be (as shown in Fig. 8.1*b*) $396 - 156 = 240$ V.

The fact that the voltage across the capacitor is greater than the total supplied may be surprising, but it is nothing to what we shall see soon!

8.2 *L, C* and *R* all in Series

Bringing R into the circuit introduces no new problem, because we have just found that L and C can always be replaced (for purposes of the calculation) by either L or C of suitable value, and the method of combining this with R was covered in the previous two chapters. Elaborating the equation given therein (Sec. 6.8) to cover the new information, we have

$$Z = \sqrt{[(X_L - X_C)^2 + R^2]}$$

If the three circuit elements are connected in the order shown in Fig. 8.2*a*, the phasor diagram, drawn according to the now familiar rules, will be as at *b*. Note that ad, representing the voltage across C, is longer than ab representing the generator voltage. If L and C were replaced by a larger C to form the equivalent circuit mentioned above, *bc* would disappear and *ad* be shortened by that amount, but *ab* would still have the same length and phase angle relative to AB.

8.3 The Series Tuned Circuit

We have already seen that the reactance of a capacitor falls and that of an inductor rises as the frequency of the current supplied to them *as a whole* to a voltage of this particular frequency, by completely removing from it both L and C. This, leaving only R, would result in the flow of a current equal to E/R.

Let us assume a voltage not unlikely in radio broadcast reception, and see what happens when $E = 5$ mV. The current at 796 kHz is then $5/10 = 0.5$ mA, and this current flows, not through R only, but through L and C as well. Each of these has a reactance of 1000 Ω at this frequency; the potential difference across each of them is therefore $0.5 \times 1000 = 500$ mV, which is just one hundred times the voltage E of the generator to which the flow of current is due.

That a simple circuit like this, without any amplifier, can enable so small a voltage to yield two such relatively large voltages, may seem incredible, but practical radio communication would not be possible without it.

8.4 Magnification

In the particular case we have discussed, the voltage across the coil (or across the capacitor) is one hundred times that of the generator. This ratio is called the *magnification* of the circuit.

We have just worked out the magnification for a particular circuit; now let us try to obtain a formula for any circuit. The magnification is equal to the voltage across the coil divided by that from the generator, which is the same as the voltage across R. If I is the current flowing through both, then the voltage across L is IX_L, and that across R is IR. So the magnification is equal to IX_L/IR, which is X_L/R or (because in the circumstances considered X_C is numerically equal to X_L) X_C/R. At any given frequency, it depends solely on L/R, the ratio of the inductance of the coil to the resistance of the circuit, or on CR. (Both of these, you may have noticed, are time constants; see Sections 3.5 and 4.8.)

To obtain high magnification of a received signal (for which the generator of Fig. 8.3 stands), it is thus desirable to keep the resistance of the circuit as low as possible.

The symbol generally used to denote this voltage magnification is Q (not to be confused with Q denoting quantity of electricity).

In any real tuned circuit the coil is not pure inductance as shown in Fig. 8.3, but contains the whole or part of R and also has some stray capacitance in parallel with it. Consequently the magnification obtained in practice may not be exactly equal to Q (as calculated by X_L/R); but the difference is usually unimportant.

8.5 Resonance Curves

At other frequencies the impedance of the circuit is greater, because in addition to R there is some net reactance. At 1250 kHz, for one particular frequency—in this case 1000 kHz—than to voltages of substantially different frequencies. The circuit is said to be *tuned* to, or to *resonate* at, 1000 kHz; and the curves are called *resonance* curves. This electrical resonance is very closely analogous to acoustical resonance: the way in which hollow spaces or pipes magnify sound of a particular pitch.

The principle on which a receiver is tuned is now beginning to appear; by adjusting the values of L or C in a circuit such as that under discussion one can make it resonate at any desired frequency. Signal voltages received from the antenna at that frequency will be amplified so much more than those of substantially different frequencies that these others will be more or less tuned out.

8.6 Selectivity

This ability to pick out signals of one frequency from all others is called *selectivity.* It is an even more valuable feature of tuned circuits than magnification. There are alternative methods (which we shall consider later) of magnifying incoming signal voltages, but by themselves they would be useless, because they fail to distinguish between a programme on the desired frequency and others.

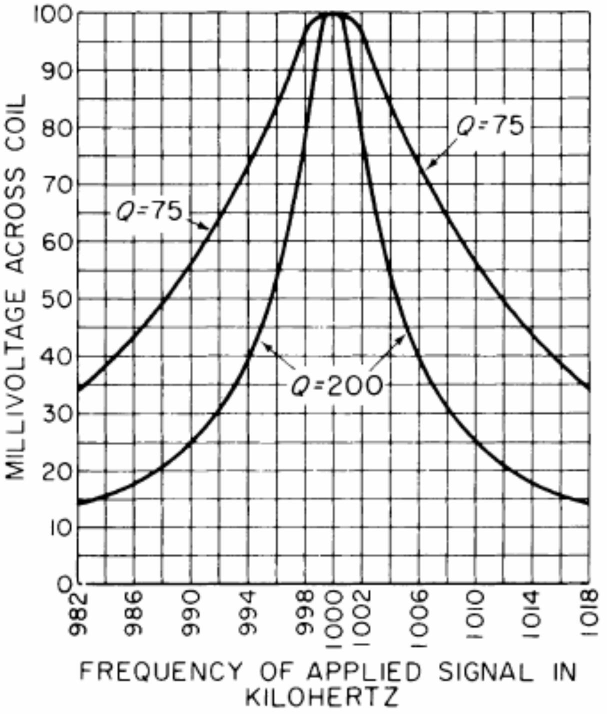


Fig. 8.6 In this diagram the curve '$Q = 200$' is the same as in Fig. 8.5, but, to enable the selectivity of the $Q = 75$ circuit to be more easily compared, the voltage injected into it has been raised from 0.5 mV to 1.33 mV, so as to make the output voltage at resonance equal to that across the $Q = 200$ circuit

The way the curves in Fig. 8.5 have been plotted focuses attention on how the value of Q affects the response at resonance. The comparative selectivity is more easily seen, however, if the curves are plotted to give the same voltage at resonance, as has been done in Fig. 8.6. This necessitates raising the input to the low-Q circuit from 0.5 V to 1.33 V. Since the maximum voltage across both the coils is now 100 mV, the scale figures can also be read as percentage of maximum. Fig. 8.6 brings out more clearly than Fig. 8.5 the better selectivity of the high-Q circuit. For a given response to a desired station working on 1 000 kHz, the 200-Q response to 990 kHz voltages is less than half that of the 75-Q circuit. In general, the rejection of frequencies well off resonance is nearly proportional to Q.

8.7 Frequency of Resonance

It is obviously important to be able to calculate the frequency at which a circuit containing known L and C resonates, or to calculate the L and C required to tune to a given frequency. The required equation follows easily from the fact that resonance takes place at the frequency which makes the reactance of the coil equal that of the capacitor:

$$2\pi f_r L = \frac{1}{2\pi f_r C}$$

where the symbol f_r is used to denote the frequency of resonance. Rearranging this, we get

$$f_r^2 = \frac{1}{(2\pi)^2 LC}, \qquad \text{so} \quad f_r = \frac{1}{2\pi\sqrt{(LC)}}$$

If L and C are respectively in henries and farads, f_r will be in hertz; if L and C are in henries and microfarads, f_r will be in kHz. But perhaps the most convenient units for radio purposes are μH and pF, and when the value of π has been filled in the result is

$$f_r \text{ (in kHz)} = \frac{159{,}155}{\sqrt{(LC)}} \qquad \text{or} \qquad f_r \text{ (in MHz)} = \frac{159.15}{\sqrt{(LC)}}$$

If the answer is preferred in terms of wavelength, we make use of the relationship $f = 3 \times 10^8/\lambda$ (Sec. 1.9) to give

$$\lambda_r = 1.885\sqrt{(LC)} \qquad (\lambda_r \text{ in m};\ L \text{ in } \mu\text{H};\ C \text{ in pF})$$

Note that f_r and λ_r depend on L multiplied by C; so in theory a coil of any inductance can be tuned to any frequency by using the appropriate capacitance. In practice the capacitance cannot be reduced indefinitely, and there are disadvantages in making it very large. A typical tuning capacitor for the medium- or low-frequency bands has a range of about 16 to 270 pF. The capacitance of the wiring and circuit components necessarily connected in parallel may add another 14 pF. This gives a ratio of maximum to minimum of $284/30 = 9.47$. But the square root sign in the above equations means that if the inductance is kept fixed the ratio of maximum to minimum f_r (or λ_r) is $\sqrt{9.47}$, or 3.075. Any band of frequencies with this range of maximum to minimum can be covered with such a capacitor, the actual frequencies in the band depending on the inductance chosen for the coil.

Suppose we wished to tune from 1620 MHz to 1620/3.075 or 527 kHz, corresponding to the range of wavelengths 185 to 569 metres. For the highest frequency or lowest wavelength the capacitance will have its minimum value of 30 pF; by filling in this value for C and 1.620 MHz for f_r, we have $1.620 = 159.15/\sqrt{(30L)}$, from which $L = 321.5\ \mu\text{H}$*.

If we calculate the value of L necessary to give 0.527 MHz with a capacitance of 284 pF, it will be the same.

In the same way the inductance needed to cover the short-wave band 9.8 to 30 MHz (30.6 to 10 metres) can be calculated, the result being 0.938 μH.

Notice how large and clumsy numbers are avoided by a suitable choice of units, but of course it is essential to use the equation having the appropriate numerical 'constant'. If in any doubt it is best to go back to first principles ($2\pi f_r L = 1/2\pi f_r C$) and use henries, farads, and Hz.

8.8 *L* and *C* in Parallel

Having studied circuits with L and C in series we come naturally to the simple parallel case, Fig. 8.7*a*. It should hardly be necessary by now to point out that the voltage *ab* is the quantity common to both L and C, and so we start with its phasor. Drawing the current phasors in the appropriate directions and assuming that in this particular example more current goes through L than through C we get the complete diagram, Fig. 8.7*b*.

If we go back to Fig. 8.1 and compare the two phasor diagrams we see that they are identical except that current and voltage have changed places. We seem to be coming across quite a lot of this sort of thing: see Secs. 4.8, 4.10 and 6.10 for example. It has a special name, but going into that just now would divert attention from our study of tuned circuits so will have to wait until Sec. 12.2.

Meanwhile we note the fact of current-voltage interchange

---

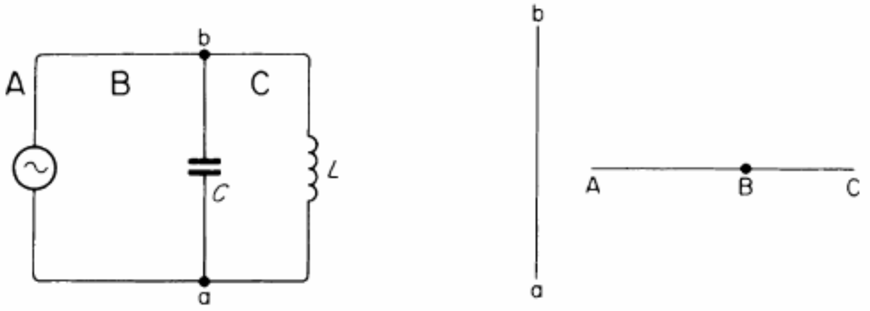
*For this reverse process it saves time to adapt the formula, thus:

$$f_r = 159.15/\sqrt{(LC)} \qquad (\mu\text{H, pF, MHz}).$$
$$\therefore f_r^2 = 159.15^2/LC$$
$$\therefore L = 25333/f_r^2 C$$

The corresponding adaptation for calculating C is

$$C = 25333/f_r^2 L$$

sufficiently to be able to grasp, without too much repetition along the lines of Sec. 8.1, that in the parallel circuit the branch *currents* are in opposite phase, so that the total current to be supplied is the *difference* between them (or their sum if one of them is taken as negative). Similarly the two *susceptances,* B_L and B_C, are opposite in sign, and this time it is the inductive one that is taken as negative. So the total susceptance in our example is negative and smaller numerically than either B_L or B_C.



(a) (b)

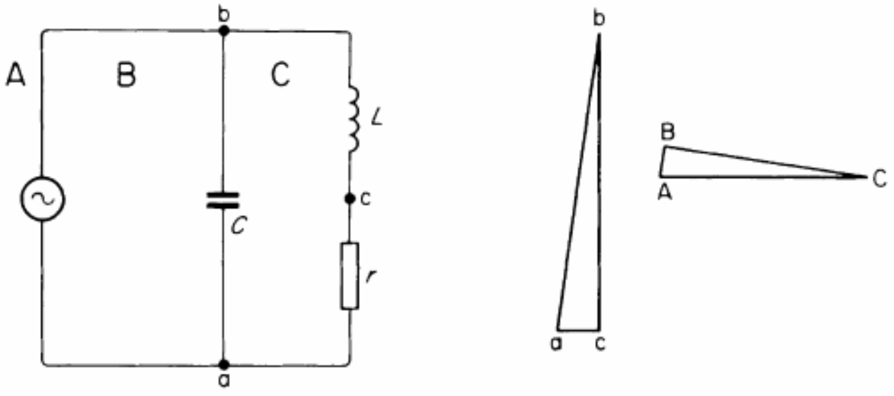
Fig. 8.7 *L* and *C* in parallel, and corresponding phasor diagrams

Suppose now that the frequency is such as to make the susceptances (and therefore the reactances) equal. The currents will therefore be equal, so that the difference between them will be zero. We then have a circuit with two parallel branches, both with currents flowing in them, and yet the current supplied by the generator is nil!

It must be admitted that such a situation is impossible, the reason being that no practical circuits are entirely devoid of resistance. But if the resistance is small, when the two currents are equal the system does behave approximately as just described. It is a significant fact, which can easily be checked by calculation of the type given in Sec. 8.7, that the frequency which makes them equal is the same as that which in a series circuit would make *L* and *C* resonate.

8.9 The Effect of Resistance

When the resistance of a parallel LC circuit has been reduced to a minimum, the resistance in the C branch is likely to be negligible compared with that in the L branch. Assuming this to be so, let us consider the effect of resistance in the L branch (Fig. 8.8), representing it by the symbol *r* as a reminder that it is small compared with the reactance. We have already studied this branch on its own (Fig. 7.4) and noted that one effect of the resistance is to make the phase angle between current and voltage less than 90°; the greater the resistance the smaller the angle. It would be quite easy to modify Fig. 8.7*b* by reducing the phase angle slightly, but to do the job thoroughly, so that if necessary we could draw the diagram to scale, let us start from scratch without reference to other diagrams and just follow the basic rules of phase relationships.



(a) (b)

Fig. 8.8 (*a*) Ideal capacitor in parallel with inductor having losses represented by resistance r in series, and (*b*) corresponding phasor diagrams

We start with the series branch made up of L and r, and as current is common to both we draw its phasor, AC, horizontally. Now we can draw ac in phase, small because r is small; and cb 90° ahead. That completes the voltage part. Having now got ab, not perfectly vertical this time, we can draw CB 90° ahead of it; let us assume it is equal to AC. That shows us the supplied current, AB. Because AC = CB, the angle ABC is less than a right angle, but as we have drawn CB at right angles to ab we deduce that although the supplied current AB looks as if it might be in phase with its e.m.f. ab it is not exactly so. It could be made so by a very slight reduction in frequency, which would increase AC and reduce CB. (This sort of information would be almost impossible to obtain from waveform diagrams, even at cost of much greater time and trouble.) In practice r is likely to be less than one-hundredth as much as X_L, so even when X_L is made exactly equal to X_C the current supplied by the generator would be small and practically in phase with its e.m.f. In other words, for most purposes it can be taken as equal to the current that would flow if the two reactive branches were replaced by one consisting of a high resistance.

8.10 Dynamic Resistance

It is of considerable practical interest to know how high this resistance is (let us call it R), and how it is related to the values of the real circuit: L, C and r. We have already seen that if $r = 0$, AB is zero, so in that case R is infinitely large. Fig. 8.8*b* shows that as r is increased, making the voltage drop across it, ac, greater, ab is swung round clockwise, and BC with it, increasing AB. This increase in current from the generator signifies a fall in R. So at least we can say that the smaller we make r in Fig. 8.8, the greater is the resistance that the circuit as a whole presents to the generator.

It can be shown mathematically that, so long as r is much less than X_L, the resistance R, to which the parallel circuit *as a whole* is equivalent at a certain frequency, is practically equal to X_L^2/r. The proof by algebra is given in textbooks, but geometry-minded readers should not find it difficult to prove from Fig. 8.8*b*.

To take an example, suppose C and L are the same as in Fig. 8.3, and r is equal to 10 Ω (Fig. 8.9*a*). At 796 kHz the reactances are each

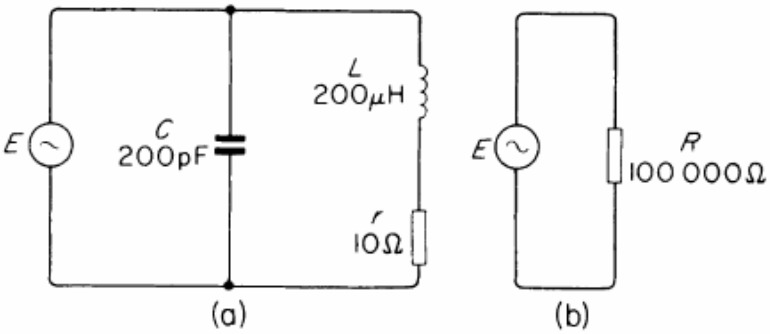


Fig. 8.9 The circuit composed of *C, L* and *r* having the values marked (*a*) can at 796 kHz be replaced by a high resistance (*b*), so far as the phase and magnitude of the current taken from the generator is concerned

1000 Ω (Fig. 8.4). So R is $1000^2/10 = 100000$ Ω (Fig. 8.9*b*). If *L, C* and r were hidden in one box and R in another, each with a pair of terminals for connecting the generator, it would be impossible to tell, by measuring the amount and phase of the current taken, which box was which. Both would appear to be resistances of 100 kΩ, *at that particular frequency*. We shall very soon consider what happens at other frequencies, but in the meantime note that the apparent or equivalent resistance, which we have been denoting by *R,* has the special name *dynamic resistance.*

8.11 Parallel Resonance

Suppose E in Fig. 8.9 is 1 V. Then under the conditions shown the current I must be 1/100 000 A, or 10 μA. Since the reactances of C and L are each 1000 Ω, and the impedance of the L branch is not appreciably increased by the presence of *r,* the branch currents are each 1000 μA. But let us now change the frequency to 1000 kHz. Fig. 8.4 shows that this makes X_C and X_L respectively 800 Ω and 1250 Ω. The current taken by C is then 1/800 A = 1250 μA and the L current is 1/1250 A = 800 μA. Since r is small, these two currents are so nearly in opposite phase that the total current is practically equal to the difference between them, 1 250 − 800 = 450 μA. This balance is on the capacitive side, so at 1000 kHz the two branches can be replaced by a capacitance having a reactance = E/I = $1/0.00045$ = 2222 Ω, or 71.6 pF. The higher the frequency the greater is the current, and the larger this apparent capacitance; at very high frequencies it is practically 200 pF—as one would expect, because the reactance of L becomes so high that hardly any current can flow in the L branch.

Using the corresponding line of argument to explore the frequencies below 1000 kHz, we again find that the total current becomes larger, and that its phase is very nearly the same as if C were removed, leaving an inductance greater than L but not much greater at very low frequencies.

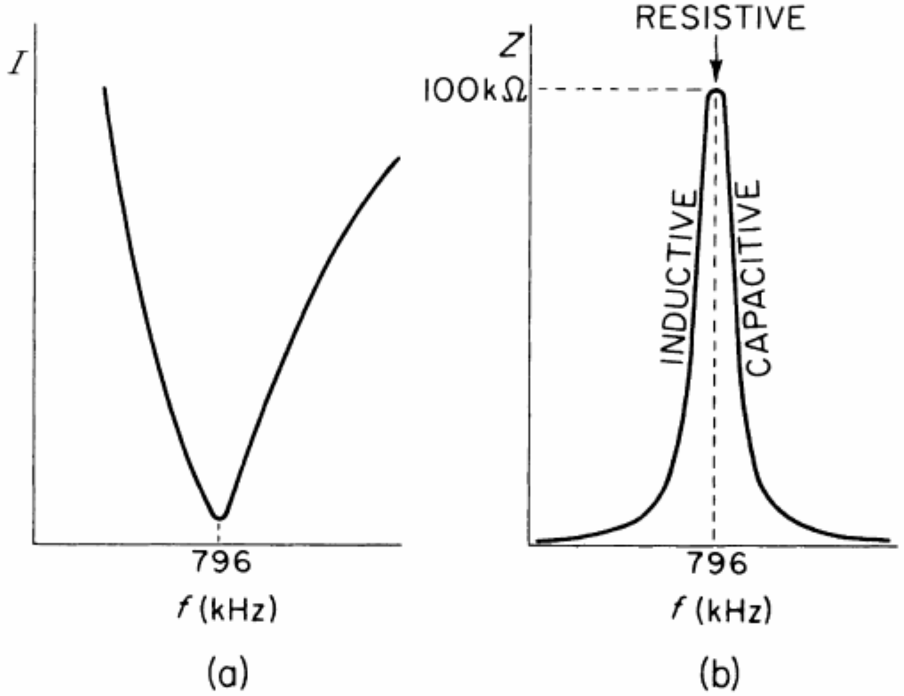


Fig. 8.10 Plotting the current and impedance of the Fig. 8.9*a* circuit as a whole at various frequencies gives the shapes shown here: (*a*) current, (*b*) impedance

Plotting the current over a range of frequency, we get the result shown in Fig. 8.10*a*, which looks somewhat like a series-circuit resonance curve turned upside down.

The graph of impedance against frequency (Fig. 8.10*b*) is still more clearly a resonance curve.

8.12 Frequency of Parallel Resonance

Before we complicated matters by bringing in r we noticed that the frequency which would make L and C resonate in a series circuit had the effect in a parallel circuit of reducing I to zero and making the impedance infinitely great. In these circumstances the frequency of parallel resonance is obviously the same as that of series resonance. But actual circuits always contain resistance, which raises some awkward questions about the frequency of resonance. In a series circuit it is the frequency that makes the two reactances equal. But in a parallel circuit this does not make the branch impedances exactly equal unless both contain the same amount of resistance, which is rarely the case. The two branch currents are not likely to be exactly equal, therefore. So are we to go on defining f_r as the frequency that makes $X_L = X_C$, or as the frequency which brings I into phase with E so that the circuit behaves as a resistance R, or as the frequency at which the current I is least?

Fortunately it happens that in normal radio practice this question is mere hair-splitting; the distinctions only become appreciable when r is a large-sized fraction of X_L; in other words, when $Q\ (= X_L/r)$ is abnormally small—say less than 5. *On the understanding that such unusual conditions are excluded, therefore,* the following relationships are so nearly true that the inaccuracy is of no practical importance. The frequency of parallel resonance (f_r) is the same as that for series resonance, and it makes the two reactances X_L and X_C numerically equal. The impedance at resonance (R) is resistive, and equal to X_L^2/r and also to X_L^2/r (Sec. 8.10). And, because $X_L = X_C$,

$$R = \frac{X_L X_C}{r} = \frac{2\pi f_r L}{2\pi f_r C r} = \frac{L}{Cr}$$

from which $$r = \frac{L}{CR}$$

In all these, r can be regarded as the whole resistance in series with L and C, irrespective of how it is distributed between them. Since $Q = X/r$,

$$R = QX, \quad r = \frac{X}{Q}, \qquad \frac{R}{X} = \frac{X}{r} \quad \text{and} \quad \frac{R}{Q} = Qr$$

All these relationships assume resonance.

8.13 Series and Parallel Resonance Compared

The resonance curve in Fig. 8.10*b* showing $R = 100$ kΩ applies to the circuit in Fig. 8.9*a*, in which $Q = X_L/r = 2\pi f L/r = 2\pi \times 796\,000 \times 0.0002/10 = 100$. If r were reduced to 5 Ω, Q would be 200, and R would be 200 kΩ. But reducing r would have hardly any effect on the circuit at frequencies well off resonance; so the main result would be to sharpen the resonance peak. Increasing r would flatten the peak. This behaviour is the same as with series resonance, except that a peak of impedance takes the place of a peak of voltage across the coil or of current from the generator. Current from the generator to the parallel circuit reaches a minimum at resonance (Fig. 8.10*a*), as does impedance of the series circuit.

At resonance a series circuit takes the maximum current from the generator, so is called an *acceptor circuit.* A parallel circuit, on the contrary, takes the minimum current so is called a *rejector circuit.* One should not be misled by these names into supposing that there is some inherent difference between the circuits; the only difference is in the way they are used. Comparing Fig. 8.9*a* with Fig. 8.3 we see the circuits are identical except for where the generator is connected. To a generator connected in one of the branches, Fig. 8.9*a* would be an acceptor circuit at the same time as it would be a rejector circuit to the generator shown.

If the generator in a circuit such as Fig. 8.9*a*, were to deliver a constant current instead of a constant e.m.f., the voltage developed across it would be proportional to the impedance of the tuned circuit and so would vary with frequency in the manner shown in Fig. 8.5. So the result is the same, whether a tuned circuit is used in series with a source of voltage or in parallel with a source of current. In later chapters we shall come across examples of both.

8.14 The Resistance of the Coil

In the sense that it cannot be measured by ordinary direct-current methods—by finding what current passes through on connecting it across a 2 V cell, for example—it is fair to describe R as a fictitious resistance. Yet it can quite readily be measured by any method suitable for measuring resistances at the frequency to which the circuit is tuned; in fact, those methods by themselves would not disclose whether the thing being measured was the dynamic resistance of a tuned circuit or the resistance of a resistor.

If it comes to that, even r is not the resistance that would be indicated by any d.c. method. That is to say, it is not merely the resistance of the wire used to make the coil. Although no other cause of resistance may appear, the value of r measured at high frequency is always greater, and may be many times greater, than the d.c. value.

One possible reason for this has been mentioned in connection with transformers (Sec. 7.10). If a solid iron core were used, it would have currents generated in it just like any secondary winding: the laws of electromagnetism make no distinction. These currents passing through the resistance of the iron represent so much loss of energy, which, as we have seen, brings the primary current more nearly into phase with the e.m.f., just as if the coil's impedance included a larger proportion of resistance. To stop these *eddy currents,* as they are called, iron cores are usually made up of thin sheets arranged so as to break up circuits along the lines of the induced e.m.f. At radio frequencies even this is not good enough, and if a magnetic core is used at all it is either iron in the form of fine dust bound together by an insulator, or a non-conducting magnetic material called a *ferrite.*

Another source of loss in magnetic cores, called *hysteresis,* is a lag in magnetic response, which shifts the phase of the primary current still more towards that of the applied e.m.f., and therefore represents resistance, or energy lost.

The warming-up of cores in which a substantial number of watts are being lost in these ways is very noticeable.

Although an iron core introduces the equivalent of resistance into the circuit, its use is worth while if it enables a larger amount of resistance to be removed as a result of fewer turns of wire being needed to give the required inductance. Another reason for using iron cores, especially in r.f. coils, is to enable the inductance to be varied, by moving the core in and out.

Even the metal composing the wire itself has e.m.fs induced in it in such a way as to be equivalent to an increase in resistance. Their distribution is such as to confine the current increasingly to the surface of the wire as the frequency is raised. This *skin effect* occurs even in a straight wire; but when wound into a coil each turn lies in the magnetic field of other turns, and the resistance is further increased by so-called *proximity effect*; so much so that using a thicker gauge of wire sometimes actually increases the r.f. resistance of a coil.

8.15 Dielectric Losses

In the circuits we have considered until now, r has been shown exclusively in the inductive branch. While it is true that the resistance of the capacitor plates and connections is usually negligible (except perhaps at very high frequencies), insulating material coming within the alternating electric field introduces resistance rather as an iron core does in the coil. It is as if the elasticity of the 'leashes' (Sec. 3.4) were accompanied by a certain amount of friction, for the extra circuit current resulting from the electron movements is not purely capacitive in phase. The result is the same as if resistance were added to the capacitance.

Such materials are described as poor dielectrics, and obviously would not be chosen for interleaving the capacitor plates. Air is an almost perfect dielectric, but even in an air capacitor a part of the field traverses solid material, especially when it is set to minimum capacitance. Apart from that, wiring and other parts of the circuit—including the tuning coil—have 'stray' capacitances; and if inefficient dielectrics are used for insulation they will be equivalent to extra resistance.

8.16 H.F. Resistance

All these causes of loss in high-frequency circuits come within the general definition of resistance (Sec. 2.18) as (electrical power dissipated) ÷ (current squared). In circuit diagrams and calculations it is convenient to bring it all together in a single symbol such as r in Fig. 8.9*a* in conjunction with a perfectly pure inductance and capacitance, L and C. But we have seen that at resonance the whole tuned circuit can also be represented as a resistance $R = L/Cr$ (Fig. 8.9*b*). We also know that a perfect L and C at resonance have an infinite impedance, so could be shown connected in parallel with R as in Fig. 8.11*a*. That they are 'invisible' to the generator is shown in the phasor diagram (*b*) by the points A and C coinciding, meaning that the current AC is zero. At the same time the capacitor and coil currents, CD and DA, are relatively large, but equal and opposite. This current diagram is unusual in being a four-sided figure with two diagonally opposite points coinciding (to put it paradoxically!). If R had been placed between C and L the current diagram would have opened out into a slim rectangle. Fig. 8.11*a* represents the actual circuit not only at resonance but (except in so far as R varies somewhat with frequency) at frequencies on each side of resonance.

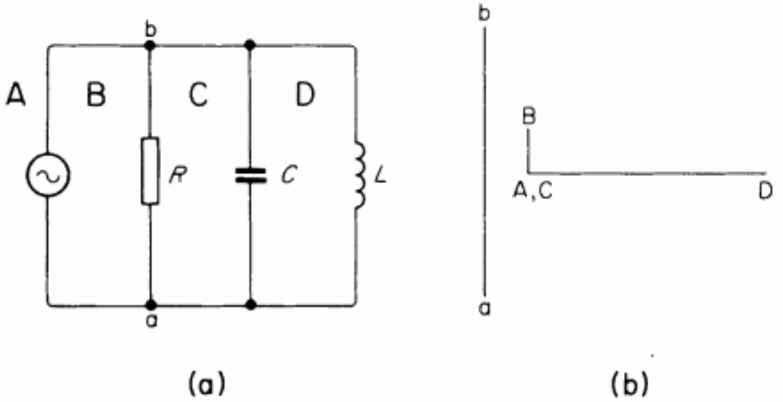


Fig. 8.11 L, C and R all in parallel, and corresponding phasor diagrams

More generally, for every resistance and reactance in series one can calculate the values of another resistance and reactance which, connected in parallel, at the same frequency, are equivalent. And if (as we are assuming) the reactance is much greater than the series resistance, it has very nearly the same value in both cases. The values of series and parallel resistance, r and R respectively, are connected by the same approximate formula $R = X^2/r$, which we have hitherto regarded as applying only at the frequency of resonance (Sec. 8.9). This ability to reckon resistance as either in parallel or in series is very useful. It may be, for example, that in an actual circuit there is resistance connected both ways. Then for simplifying calculations it can be represented by a single resistance, either series or parallel.

The capacitance required to tune the coil to 300 kHz, or 0.3 MHz, is (Sec. 8.7, footnote) $C = 25\,333/f_r^2\,L = 25\,333/(0.3^3 \times 1\,200) =$ 235 pF. This, of course, includes the tuning capacitor itself plus the stray capacitance of the wiring and of the coil. The whole of this is equivalent to a perfect 235 pF in parallel with a resistance of, shall we say, 0.75 MΩ (Fig. 8.12*c*). The reactance, X_C, is bound to be same as X_L at resonance; but to check it we can work it out from $1/2\pi fC = 1/(2\pi \times 0.3 \times 0.000\,235) = 10^6/(2\pi \times 0.3 \times 235) =$ 2260 Ω. Q_C is therefore $R_C/X_C = 750\,000/1260 = 332$.

If the two components are united as a tuned circuit, the total loss can be expressed as a parallel resistance, R, by reckoning R_L and R_C in parallel: $R_L R_C/(R_L + R_C) = 0.205 \times 0.75/0.955 = 0.161$ MΩ (Fig. 8.12*d*). Although this is less than R_L, it indicates a greater loss. The Q of the tuned circuit as a whole is $R/X = 161\,000/2260 = 71$. Incidentally, it may be seen that this is $Q_L Q_C/(Q_L + Q_C)$. The total loss of the circuit can also be expressed as a series resistance, $r = X^2/R = 2260^2/161\,000 = 31.8\ \Omega$ (Fig. 8.12*e*). Calculating Q from this, $X/r = 2260/31.8$, gives 71 as before.

Before going on, the reader would do well to continue calculations of this kind with various assumed values, taking note of any general conclusions that arise. For instance, a circuit to tune to the same frequency could be made with less inductance; assuming the same Q, would the dynamic resistance be lower or higher? Would an added resistance in series with the lower-L coil have less or more effect on its Q? And what about a resistance in parallel?

8.17 Cavity Resonators

As radio communication developed, carrier frequencies became higher, increasing from less than 1 MHz in the early days of broadcasting to thousands of MHz in modern radio links. The inductance and capacitance used in tuned circuits have correspondingly decreased; for example a long-wave tuning coil might have 250 turns, a medium-wave one 80 turns and a coil for 100 MHz about five turns. A coil with less than one turn is difficult to envisage and, to resonate at hundreds of MHz, a different type of tuning system is adopted using transmission lines (Sec. 16.10). To achieve resonance at frequencies above approximately 2000 MHz even this technique has to be abandoned and the following system is used.

Imagine a conventional LC tuned circuit in which, to achieve a high resonance frequency, the inductor has been reduced to a single loop of wire and the capacitor to two small parallel plates. Such a tuned circuit, illustrated in Fig. 8.13*a*, resembles that used by Hertz in his early experiments on radio, and might achieve a frequency in the low hundreds of MHz. To raise the frequency we must reduce

CHAPTER 9

Diodes

9.1 Electronic Devices

We now have all the essential basic principles of circuits, and it is time to strike out along a new line—electronics. Although all the electric currents we have been discussing in the last seven chapters consisted of electrons in motion there has been no real need to consider them as such. It would have been just about as easy to deal with the subject using one of the theories of the electric current in vogue before electrons were discovered. But now we are going to study electronic devices, which could not be satisfactorily explained without reference to electrons. It is these devices that are responsible for the rapid and multifarious developments included under the broad name of electronics.

There are two main divisions of electronic devices: those in which the currents pass through vacuum or gas, and those in which they pass through semiconducting solids. The first includes those called valves in Britain and tubes in America, and the second includes transistors.

Although some simple semiconductor devices were used, without knowing how they worked, for several decades before 1948, it was not until that year, when the transistor was invented as a result of research into the electronic nature of semiconductors, that they began to overtake thermionic valves (as they are named more fully), which are reckoned to have started in 1904.

Electronic devices can be, and often are, classified according to the number of their *electrodes.* An electrode is not very easy to define precisely, but roughly it is any part of the device where a current leaves or enters or by means of which it is electrically controlled. Often each electrode has a terminal or wire by which connection is made to a circuit, but sometimes electrodes are connected together inside the device.

9.2 Diodes

The least number of electrodes is two, and devices so constituted are called *diodes.* Although they are relatively limited in what they can do—their chief capability is allowing currents to flow only one way—they embody most of the basic principles needed for understanding all electronic devices. So as not to be too theoretically minded, however, we may at once note two most important practical uses of diodes. Working at low (or power) frequencies such as 50 Hz, they enable a.c. to be converted to d.c. such as is required for running nearly all electronic equipment. This function is called *rectification.* And by rectifying radio-frequency currents, diodes enable radio signals to be made evident, as mentioned in Sec. 1.11. This function, although basically rectification, is often distinguished as *detection* and is the subject of Chapter 18.

Although thermionic diodes have been superseded almost entirely by semiconductor (or solid-state) types, their principles are easier to understand and are a necessary introduction to other valves and to cathode-ray tubes. So we begin with them.

9.3 Thermionic Emission of Electrons

So far we have considered an electric current as a stream of electrons along a conductor. The conductor is needed to provide a supply of 'loose' or mobile electrons, ready to be set in motion by an e.m.f. An e.m.f. applied to an insulator causes no appreciable current because the number of mobile electrons in it is negligible.

To control the current it is necessary to move the conductor, as is done in a switch or rheostat (a variable resistor). This is all right occasionally, but quite out of the question when (as in radio) we want to vary currents millions of times a second. The difficulty is the massiveness of the conductor; it is mechanically impracticable to move it at such a rate.

This difficulty can be overcome by releasing electrons, which are inconceivably light, from the relatively heavy metal conductor. Although in constant random agitation in the metal, they have not enough energy at ordinary temperature to escape beyond its surface. But if the metal is heated sufficiently they 'boil off' from it. A source of free electrons such as this is called a *cathode.* To prevent the electronic current from being hindered by the surrounding air, the space in which it is to flow is enclosed in a glass bulb and as much as possible of the air pumped out, giving a vacuum. We are now well on the way to manufacturing a thermionic valve.

Cathodes are of two main types: *directly heated* and *indirectly heated.* The first, known as a *filament* and now seldom used, consists of a wire heated by passing a current through it. To minimise the current needed, the wire is usually coated with a special material that emits electrons freely at the lowest possible temperature. The indirectly heated cathode is a very narrow tube, usually of nickel, coated with the emitting material and heated by a separate filament, called the *heater*, threaded through it. Since the cathode is insulated from the heater, three connections are necessary compared with the two that suffice when the filament serves also as the source of electrons, unless two are joined together internally. In either case only one connection counts an an electrode—the cathode; that it is heated electrically is only for convenience. So except in Fig. 9.1 the heating arrangement will be omitted.

One advantage of indirectly heated cathodes is that they can be heated by a.c., as explained in Sec. 27.12; another is that the whole cathode is at the same potential.

The electrons released by the hot cathode do not tend to move in any particular direction unless urged by an electic field. Without such a field they accumulate in the space around the cathode. Because this accumulation consists of electrons, it is a negative charge, known as the *space charge,* which repels new arrivals back again to the cathode, so preventing any further build-up of electrons in the vacuous space.

9.4 The Vacuum Diode Valve

To overcome the stoppage caused by the negative space charge it is necessary to apply a positive potential. This is introduced into the valve by means of the second electrode, a metal plate or cylinder called the *anode.*

When the anode is made positive relative to the cathode, it attracts electrons from the space charge, causing its repulsion to diminish so that more electrons come forward from the cathode. In this way current can flow through the valve and round a circuit such as in Fig. 9.1. But if the battery is reversed, so that the anode is more negative than the cathode, the electrons are repelled towards their source, and no current flows. The valve therefore permits current to flow through it in one direction only, like the air valve attached to a tyre, and this is why it was so named. The anode battery or other supply is often called the h.t. (signifying 'high tension') and the filament or heater source the l.t. ('low tension').

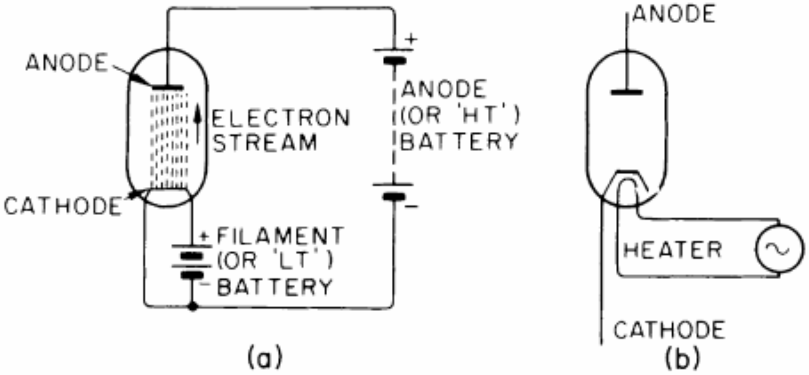


Fig. 9.1 (*a*) Electron flow from cathode to anode in a directly-heated diode. The alternative indirect method of heating is shown at (*b*)

If the anode of a diode is slowly made more and more positive with respect to the cathode, as for example by moving the slider of the a time the replacement of thermionic diodes by semiconductor ones, both for power and signals (rectifiers and detectors). So we go on now to study these.

9.5 Semiconductors

In Sec. 2.3 solid materials were grouped into conductors and insulators. The difference was explained by supposing that conductors are substances in which the electrons belonging to their atoms are free to move, under the influence of an electric field, in one direction (towards the positive pole) and so to create an electric current, whereas the electrons of insulators are, as it were, held on elastic leashes so that an electric field (unless extremely strong) is unable to move them far from the home atoms. This is a very much over-simplified view of the matter, but it will do as a first approach. For a more advanced but still not highly mathematical treatment, see the author's *The Electron in Electronics.*

Not all the electrons in atoms are involved, but only the outermost ones, known as *valence* electrons. This term first arose in chemistry, because it happens that the chemical behaviour of substances depends mainly on the number and arrangement of these outer electrons. There are never more than seven of them per atom, though the normal total number of electrons per atom, called the *atomic number*, can be anything up to about 100.

Substances in which all the atoms have the same atomic number are called *elements.* The semiconductor materials most important in electronics are the elements silicon and germanium, which are tetravalent (meaning that they have four valence electrons per atom). Their atomic numbers are 14 and 32 respectively, but for our purpose all except the four valence electrons may be lumped together with the remainder of the atoms; that is to say, the positively charged nucleus. Around this relatively heavy main body of the atom the valence electrons can be visualised as satellites (Fig. 9.4). The positive charge on any atom is exactly neutralised by the whole normal complement of electrons, so if four electrons are subtracted there must be a positive charge of four electron units, as shown.

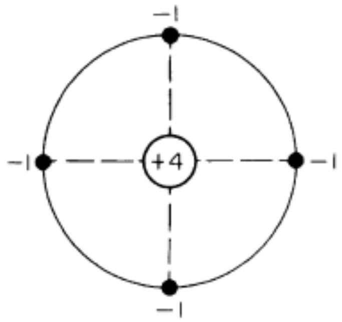


Fig. 9.4 The four valence electrons in a tetravalent atom can be visualised as satellites around the rest of the atom. These parts are electrically charged as shown

Now it happens that the full capacity of the outermost or valence orbit is eight electrons. Atoms show a general tendency to fill up vacancies in such orbits in some way or other, and one of these ways is the creation of what are called *covalent bonds.* If each atom shares each of its four valence electrons with another atom, so that both atoms have two electrons in common, every atom has eight electrons in its valence orbit and is attached to four other atoms. These are arranged symmetrically in three dimensions, so a model would be needed to represent the system properly; but Fig. 9.5 shows the result of squashing a bit of such a model flat so as to get it on a sheet of paper. Because every atom is identical with every other, the structure so formed is made up of atoms in perfectly regular array, like a case of tennis balls packed for export. But we must remember that the atoms are inconceivably small and numerous. Broken-off pieces of such an atomic structure naturally tend to have flat faces at certain definite angles to one another, and this description fits what we call crystals.

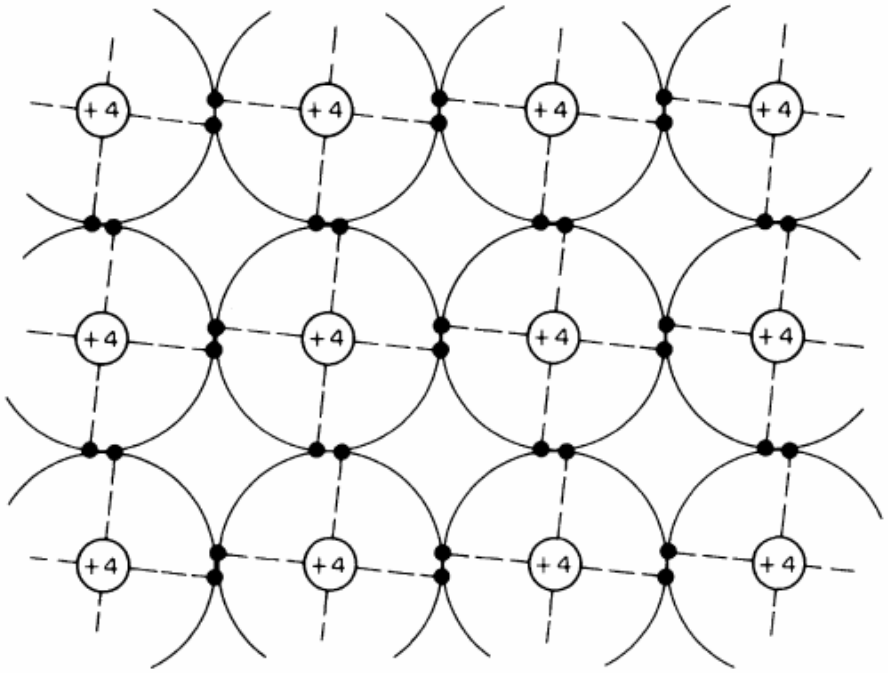


Fig. 9.5 Diagrammatic representation of the four-fold linkages of tetravalent atoms such as silicon

9.6 Holes

A crystal of this kind, if perfect, would have no electrons free to roam around and would therefore be a perfect insulator. And in fact one of the tetravalent elements, carbon, crystallises in this way (though not very readily!) as diamond, which is a very good but rather expensive insulator. But the atoms of silicon—and even more likely, to an e.m.f.) the resulting current is made up of electrons moving towards the positive end and holes towards the negative end. Owing to the more complicated way in which holes 'move', their mobility is less than that of electrons. This means that they move slower.

Just as the only things that really moved in Fig. 9.6 were tiles, so the only things that move in conduction by holes are electrons. Teachers are sometimes so anxious to emphasise that holes are not real positive charges that there may be difficulty later on when certain experiments show that holes really do behave like positive charges. Although not strictly scientifically, Fig. 9.6 covers this point too. We can all agree that both kinds of tile current are brought about solely by movements of tiles. That corresponds to the statement that a hole movement one way is really an electron movement the opposite way. But looking at the two 'charge carriers' (often called 'current carriers') we must also admit that one is a hump and the other a hollow. As these irregularities stand for negative and positive polarities respectively, one kind of current is a negative charge moving one way, and the other kind is a positive charge moving the other way. So although the causes are the same the results are different. That corresponds to the statement that a hole behaves physically as a positive charge.

9.7 Intrinsic Conduction

Having grasped the significance of holes, we may perhaps gain further insight into this kind of conduction by considering a vast array of small balls on a flat tray. They are kept in their places by resting in shallow depressions ('holes') in the tray. A difference of potental between the ends of the tray is represented by raising one end slightly, the raised end being 'negative'. Such a tilt would cause no movement of balls ('electrons'), except perhaps a slight lurch to the side of each hole nearest the 'positive' end, representing a displacement or capacitive current (Sec. 3.8). This is like a crystal at a very low temperature, when it is an insulator.

The effect of raising the temperature could be shown by making the tray vibrate with increasing vigour. At a certain intensity a ball here and there would be shaken loose, and on a perfectly level tray (no e.m.f.) would wander around at random. In doing so it would sooner or later fall into a hole—not necessarily the one from which it came. There would therefore be some random 'movement' of holes. At any given degree of vibration (temperature) a balance would eventually be reached between the rate of shaking loose and the rate of re-occupation (called in electronics recombination), but the greater the vibration the greater the number of free electrons and holes at any one time, and the greater the drift of balls towards the 'positive' end of the tray when tilted. One result of this movement would be a tendency for more holes to be reoccupied at the lower end than the upper; in effect, there would be a movement of holes up the tray towards the negative end.

To represent a sustained e.m.f. it would be necessary for the balls that reached the bottom of the tray to be conveyed by the source of the e.m.f. to the top end, where their filling of the holes would be equivalent to a withdrawal of the holes there.

If you have been able to visualise this model sufficiently you should by now have gathered that at very low temperatures a semiconductor crystal is an insulator, but that as the temperature rises it increasingly conducts. In other words its resistance falls, in contrast to that of metals, which rises with temperature. This sort of conduction is called *intrinsic*, because it is a property of the material itself.

A similar effect is caused by light; a fact that is turned to advantage in the use of semiconductors as light detectors and measurers.

Unless a substance crystallises under ideal condition, it tends to be a jumble of small crystals, at the boundaries of which there are numerous breaks in the regularity of the lattice, releasing space electrons; this is yet another cause of conduction.

Even although conduction due to these causes is contributed to by both electrons and holes, the combined number falls far short of the free electrons in metals, so the result is only 'semiconduction'. In fact, as we shall see later, to be suitable material for most electronic devices the intrinsic conduction at working temperatures should be as nearly as possible negligible. It is because silicon fulfils this condition better than germanium that it has largely superseded germanium, especially for use at high temperatures.

9.8 Effects of Impurities

So far we have assumed that our crystal, however it may be disturbed by heat, light, or its original formation, is at least all of one material. In practice however nothing is ever perfectly pure. Now if the impurities include any element having five valence electrons (such as phosphorus, arsenic or antimony) its atoms will be misfits. When they take their places in the crystal lattice—as they readily do—one electron in each atom will be at a loose end, more or less free to wander and therefore to conduct. Such an incident is represented in Fig. 9.7. Intruding atoms that yield spare electrons are called *donors*. It might be supposed that because each surplus electron leaves behind it an equal positive charge it thereby creates a hole, but it should be understood that the technical term 'hole' applies only to a *mobile* positive charge. In this case all four valence electrons needed for the lattice are in fact present, so there is no vacancy into

9.9 P-N Junctions

We can have, then, p-type material, in which conduction is by holes (positive carriers); n-type, in which it is by electrons (negative carriers); and what is sometimes called i-type (for 'intrinsic'), in which it is by both in equal numbers. None of these by itself is particularly useful; it is when more than one kind are in contact that interesting things begin to happen. Just bringing two pieces together is not good enough, however; the combination has to be in one crystalline piece, divided into two or more regions having different impurities. Various manufacturing methods have been devised for achieving this. The process of introducing a desired proportion of p or n impurity is called *doping*.

Fig. 9.10 shows with standard symbols a small section of such a combination at its boundary between p and n materials, both equally doped let us asume. The boundary is called a p-n *junction*. Compare it with Fig. 9.9 where the two materials are shown separately. Let us suppose first that there is no electric field to influence matters. But at room temperature, which is nearly 300°C above absolute zero (−273°C), the spare electrons on the n side and holes on the p side are in continuous random motion, like the intrinsic ones illustrated by the tray experiment. This motion tends to make any concentration of them gradually spread out into a uniform distribution—a process known as *diffusion*. It is by this process that when a bottle of ether is unstoppered in a room the smell pervades the whole room within seconds. Unlike intrinsic mobile charges, which for the moment we are ignoring, and impurity mobile charges within their own separate zones in Fig. 9.9, all of which from the start are distributed uniformly throughout the material because that is how they are generated, the holes have the entirely holeless n region in front of them. So they begin to diffuse across the junction. Directly they do so they encounter electrons diffusing in the opposite direction and combine with them, eliminating both so far as our diagram (restricted to electric charges) is concerned. The same story applies in reverse to the electrons. Because of this recombination process the mobile charges are prevented from diffusing far into the opposite

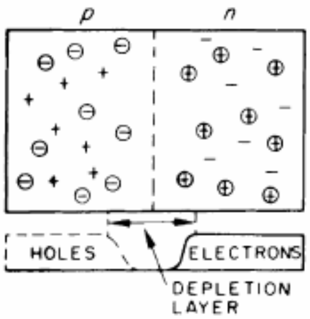


Fig. 9.10 Where p and n regions adjoin in the same crystal, diffusion across the junction and reciprocal cancellation of mobile charges creates a depletion layer. Here the fixed charges set up a potential difference that stops further diffusion

Suppose however that the e.m.f. is applied the other way round, as in Fig. 9.12, and is increased gradually from zero. The polarity of this forward bias, as it is called, is now such as to repel electrons and holes *towards* the junction, and the first effect is to neutralise some of the fixed charges there and reduce the potential barrier. This movement constitutes only a small temporary 'capacitive' current. But when the applied e.m.f. is sufficient to exceed the initial barrier p.d. the carriers are able to flow across the junction. Because of recombinations when the electrons and holes meet, the flow of holes to the right begins to die out even before the junction is reached, and has done so completely a short distance beyond it. So current through the semiconductor, except close to the junction, is by holes on the p side and by electrons on the n side, as indicated at the foot of Fig. 9.12. The fact that near the junction there are some carriers of the 'wrong' polarity will be found later to be very important. Their number can be greatly increased if one region is doped more heavily than the other, instead of equally as we have been assuming.

If we remember that there are no positive charge carriers in metals, we may wonder where the supply of holes comes from to replenish the p zone as they are drawn to the right and are neutralised by electrons. They are in fact created by withdrawal of electrons at the positive terminal of the semiconductor.

9.10 The Semiconductor Diode

Summing up, a p-n junction allows current to flow one way (provided that the e.m.f. applied is not too small) but not the other.

In short, it is a rectifier. The foregoing explanation of why it rectifies may seem complicated in comparison with the action of the vacuum diode (Sec. 9.4), so let us look at them side by side, as in Fig. 9.13. Here diodes of both vacuum and p-n junction types are connected both ways across a battery. At *a* we are reminded that current can consist either of negative charges flowing to + or positive charges to –. Or, of course, both at once. When a vacuum diode is connected as at *b*, current can flow, because the heated

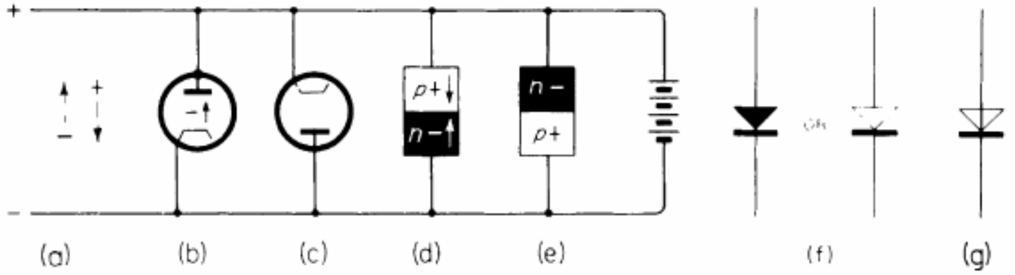


Fig. 9.13 Showing how in diodes the presence or absence of current depends on the presence or absence of the right kind of charge carriers. (*g*) is the general symbol for a diode

because of the attraction of the positive space charge there, and electrons released on the n side with sufficient force to propel them across the barrier. Hole flow is balanced in the same way. So there is no net current. But once enough reverse bias has been applied to raise the barrier too high for any majority carriers to get over it, the only current flow consists of minority carriers only, and is in the direction of the bias.

The behaviour of this leakage current, in reaching a saturation value unaffected by further increase in voltage but steeply increasing with rise in temperature, may remind us of the temperature-limited vacuum diode (Sec. 9.4) but the mechanism of the effect is quite different.

9.11 Diode Characteristics

We can now account for the features of the current/voltage characteristics of a typical silicon junction diode, Fig. 9.14. To positive voltages below about 0.6 V the response is very small, as they are occupied mainly in dismantling the potential barrier at the junction. Above that level of applied voltage the barrier is down and the forward current increases steeply; soon it is limited chiefly by the resistance of the semiconductor material and external circuit. Compared with such a large impurity current the intrinsic current at ordinary temperatures is small, so the effect of a moderate change in

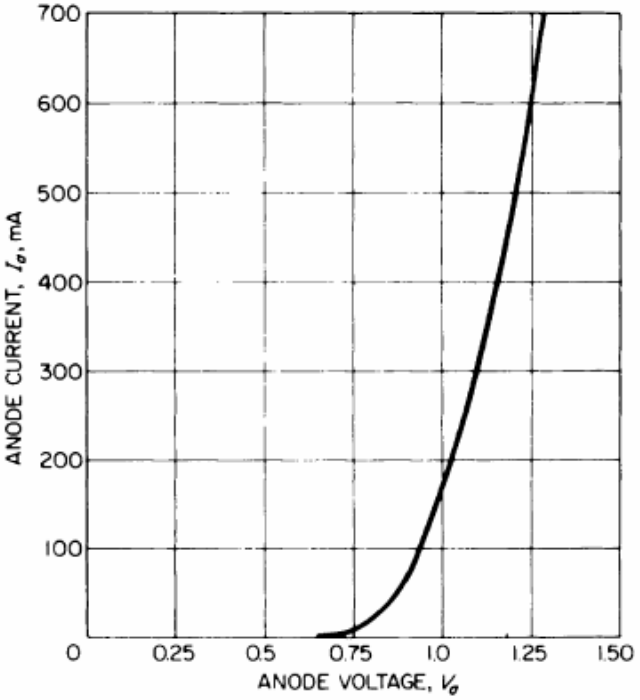


Fig. 9.14 Current/voltage characteristics of a typical small-signal silicon junction diode

protection against vibration. Here the diode element is sandwiched between two plates, held rigidly by the glass envelope.

A wide variety of characteristics are obtainable by varying the impurity in the material, and other details. For example see Sec. 14.8. The capacitance of a silicon diode, undesirable for signal rectification, can be made use of in voltage-varied capacitors, or *varactors.* The capacitance of a diode, being dependent on the thickness of the depletion layer, and that thickness varying with the reverse voltage, as shown by comparing Fig. 9.11 with 9.10, the capacitance must vary with reverse voltage. Since increasing the voltage increases the thickness of the depletion layer, it reduces the capacitance (Sec. 3.3). This property has uses such as tuning receivers (Sec. 21.6).

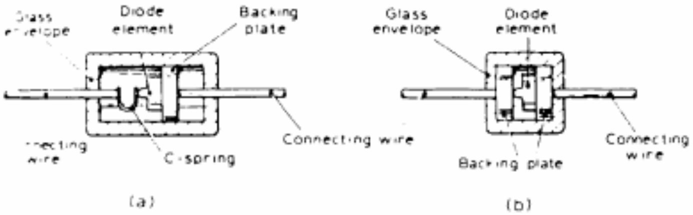


Fig. 9.15 Construction of a small-signal silicon junction diode. (*a*) spring-contact, (*b*) 'whiskerless'

One quite essential thing we have been taking for granted is the possibility of connecting the p and n regions of a diode to an external circuit. This routine necessity is actually too difficult a subject to discuss in this book, and the science of it is still not fully understood, so manufacturers have to rely mainly on practical experience. The problem is to make a joint between the semiconductor crystal and metal wire, that does not rectify on its own. The ideal is a connection that has no resistance at all; that being impossible, the aim is a resistance that is low and linear (ohmic).

9.12 Recapitulation

Even an elementary introduction to semiconductors is not easy to grasp at once, and as it is important to have at least the main points clear before going any farther, let us recapitulate. To avoid a lot of 'ifs' and 'buts' let it be assumed that the basic semiconductor material is silicon ('Si') or germanium ('Ge') and that exceptional conditions are excluded.

At low temperatures, perfectly pure material would be an insulator, because the active or valence electrons would all be occupied in holding the atoms together in regular crystalline array and there would be none to spare as charge carriers. At slightly below room temperatures (Si) or well below (Ge) the heat energy is sufficient to ionise enough atoms (i.e., release electrons from them)

CHAPTER 10

Triodes

10.1 The Vacuum Triode Valve

Valuable though diodes are, radio and electronics would not have got very far on only two electrodes. The important date is 1906, when De Forest tried adding a third electrode as a sort of tap to control the flow of current between cathode and anode. It is called the *grid,* and usually takes the form of an open spiral of thin wire wound closely around the cathode so that in order to reach the anode the electrons from the cathode have to pass between the turns of wire.

If the potential of the grid is made positive with respect to the cathode it will assist the anode to neutralise the space charge of electrons around the cathode, so increasing the anode current; and, being nearer to the cathode, one grid volt is more effective than one anode volt. If, on the other hand, it is made negative it will assist the space charge in repelling electrons back towards the cathode.

Fig. 10.1 shows the apparatus needed to take characteristic curves of this three-electrode valve, or *triode.* And in Fig. 10.2 are some such curves, showing the effects of both anode and grid voltages. (These curves are somewhat idealised, being straighter and more parallel than with real valves.) Each of them was taken with the fixed grid voltage indicated alongside. Notice that this voltage, like all others relating to a valve, is reckoned from the cathode as zero. If, therefore, the cathode of a valve is made two volts positive with respect to earth, while the grid is connected to earth, it is correct to describe the grid as 'two volts negative', the words 'with respect to the cathode' being understood. In a valve directly heated by d.c., voltages are reckoned from the negative end of the filament.

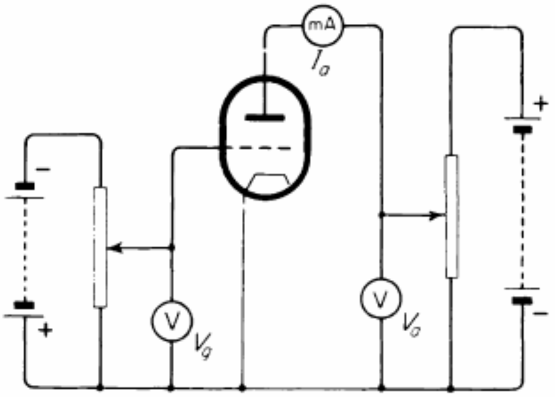


Fig. 10.1 Circuit for taking characteristic curves of a thermionic triode

10.2 Amplification Factor

One would expect that if the grid were made neither positive nor negative the triode would be much the same as if the grid were not there; in other words, as if it were a diode. This guess is confirmed by the curve in Fig. 10.2 marked '$V_g = 0$', for it might easily be a diode curve. Except for a progressive shift towards the right as the grid is made more negative, the others are almost identical. This means that while a negative grid voltage reduces the anode current in the way described, this reduction can be counterbalanced by a suitable increase in anode voltage. In the valve for which curves are shown, an anode current of 40 mA can be produced by an anode voltage of 120 if the grid is held at zero potential. This is indicated by the point A. If the grid is now made 6 V negative the current drops to 16 mA (point B), but can be brought up again to its original value by increasing the anode voltage to 180 V (point C).

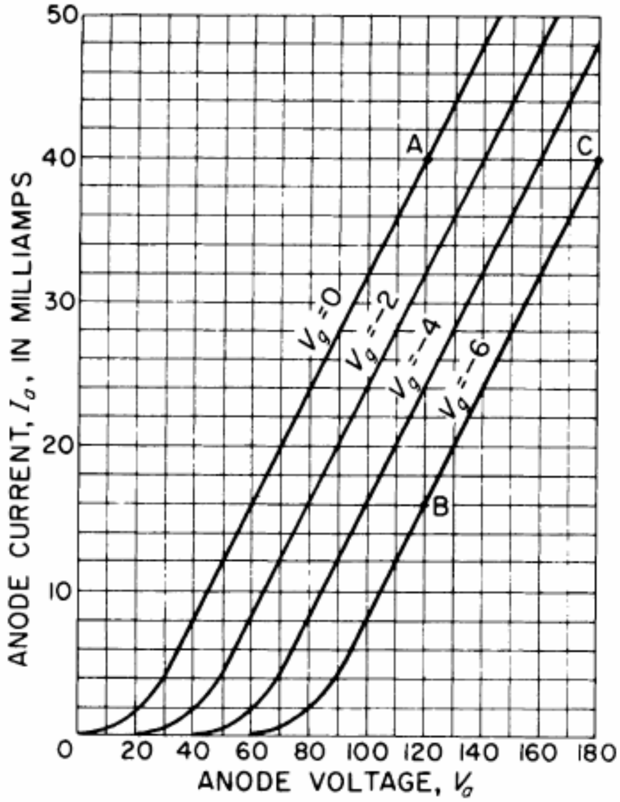


Fig. 10.2 Characteristic curves of triode valve, at different grid voltages

Looked at another way, the distance A to C represents 60 V on the V_a scale and –6 V on the V_g scale, with no change in I_a. So we can say that a change of 6 V at the grid can be compensated for by a change of 60 V, or ten times as much, at the anode. For reasons that will soon appear, this ratio of 10 to 1 is called the *amplification factor* of the valve. It is yet another thing to be denoted by μ.

10.3 Mutual Conductance

Another measure of the control that the grid has over the electron stream in a valve is the change of anode current that results from a given change in grid voltage, the anode voltage meanwhile remaining constant. Comparing points A and B in Fig. 10.2 we note that increasing the grid voltage by 6 V increases the anode current by 24 mA. In Sec. 2.12 the number of amps made to flow, per volt, was the definition of the siemens, the basic unit of conductance. Here now we have a current change brought about in a circuit (the anode circuit) by a voltage somewhere else (between cathode and grid). To distinguish this conductance from the ordinary sort it is called *mutual conductance* and given the special symbol g_m. Although its value in this example would correctly be stated as 4 mS (4 millisiemens) one is more likely to come across it in milliamps per volt. Mutual conductance is known as *transconductance* in the USA and is commonly called 'slope' as later in this chapter.

To find the value of g_m from an I_a/V_a graph, as we have just done, it must have curves for at least two different grid voltages. A more suitable form of characteristic curve for this purpose is obtained by plotting I_a against V_g, taking care to keep V_a constant, as in Fig. 10.3.

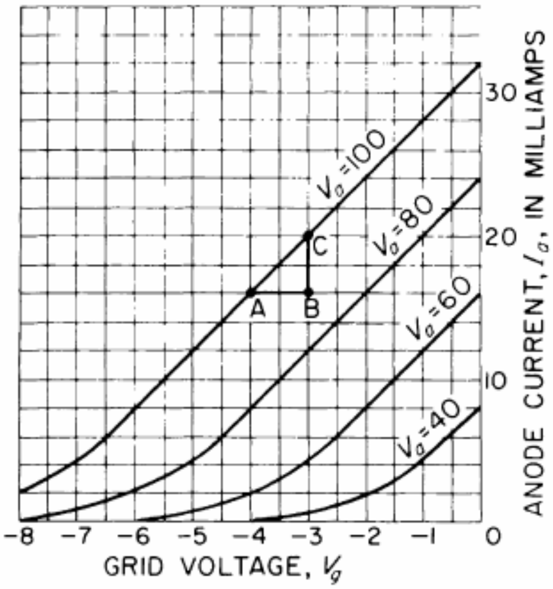


Fig. 10.3 Four samples of anode-current/grid-voltage characteristic curve, each taken at a different anode voltage

A single curve of this kind is sufficient to show g_m. For instance BC represents the increase in I_a caused by an increase in V_g represented by AB; so g_m is given by BC/AB—in this case 4 mA/V, as before. Although they can be drawn by direct measurement, using the Fig. 10.1 set-up, the curves in Fig. 10.3 were actually derived from those in Fig. 10.2, by noting (for example) the anode currents corresponding to the four grid voltages where they cut the vertical $V_a = 100$ line; and so on for the other three values of V_a.

The usefulness of identifying g_m with the slope of such curves is that one can see at once that it is not everywhere the same. Near the foot of each curve the slope, and therefore g_m, is less. And if we examine genuine valve curves we find that they are not so straight and parallel as these, so that g_m varies to some extent even along the upper stretches. Thus, although g_m used sometimes to be called a valve constant, in actual fact it is not at all constant, so is better referred to as a *parameter* of the valve. The other parameter we have encountered, μ, also varies to some extent with V_a and V_g, but much less so than g_m does.

If a characteristic curve really is appreciably curved (technically, non-linear) then not only does its slope vary from point to point but even at a given point it depends on the amplitude of the signal. In fact, slope hardly has a meaning unless the signal amplitude is small enough for the part of the curve used by it to be practically straight (linear), so a.c. parameters are often called small-signal parameters.

10.4 Anode Resistance

By this time you may be wondering if the slope of the curves in Fig. 10.2 has any special significance and name. Clearly it is a ratio of change in I_a caused by a change in V_a, V_g being kept constant. So it too is a conductance, and this time a straightforward one, because both current and voltage refer to the same thing: the cathode-to-anode path through the valve. This slope is therefore called anode conductance, and its symbol is g_a. More often, however, it is turned upside down, $1/g_a$, and this reciprocal of g_a is, as one would expect, called *anode resistance* and denoted by r_a.

When considering the diode curve, Fig. 9.3, we calculated a kind of resistance in the same way as we did with Ohm's law, as V_a/I_a, but there was not very much point in doing so, because, as we saw then, a valve is not a proper resistance and does not conform to Ohm's law. We shall see later that in most uses of valves one is more interested in *changes* in I_a, V_g, etc., than in their values relative to zero, so that is why in regard to triodes we have been working in terms of changes or differences. It represents what a valve does with an alternating voltage or current superimposed on the d.v. or d.c. corresponding to a selected point on the valve curve. So r_a is more precisely termed *anode a.c. resistance.* In Fig. 10.2 it would be found as AC/BA = 60/24 = 2.5kΩ (kΩ instead of Ω because BA is in mA). And of course g_a, being $1/r_a$, is $1/1.25$ = 0.4mA/V or mS.

If we have been keeping the figures for this example in mind we will remember that g_m is 4 mA/V, which is now seen to be 10 times g_a. And μ is 10. This is not just a coincidence. Because V_g has μ times the effect on I_a that V_a has, and μ is the ratio of their effectivenesses in this respect, it must always be true (provided all three parameters refer to the same valve under the same conditions) that

$$g_m = \mu g_a$$

In terms of r_a, therefore

$$g_m = \frac{\mu}{r_a}, \quad \text{or} \quad \mu = g_m r_a, \quad \text{or} \quad r_a = \frac{\mu}{g_m}$$

We see, then, that if any two of these three parameters are known they are all known.

10.5 Alternating Voltage at the Grid

In Fig. 10.4 we have an I_a/V_g curve for a typical triode. Suppose that, as suggested in the circuit diagram, we apply a small alternating voltage, v_g, to the grid of the valve, what will the anode current do?

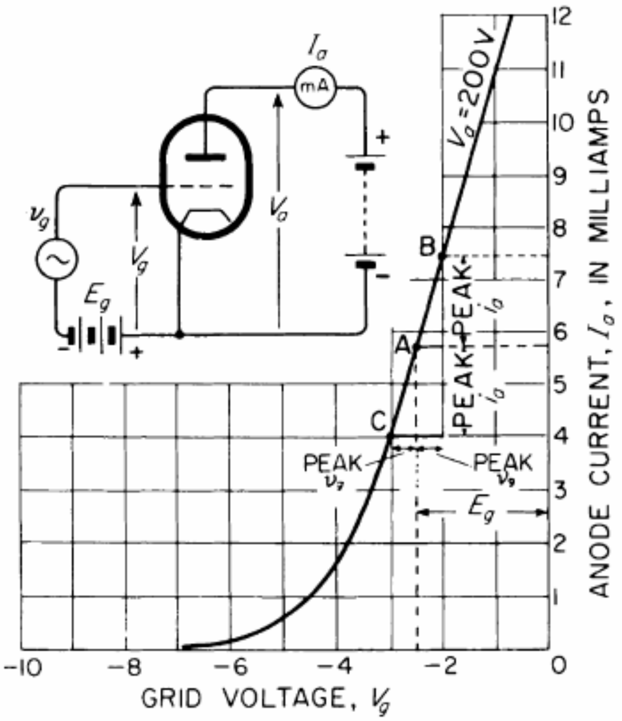


Fig. 10.4 The result of applying an alternating e.m.f. (v_g) to the grid of a triode can be worked out from the I_a/V_g curve, as here. The fixed starting voltage, E_g, in this case is −2.5 V

If the batteries supplying anode and grid give 200 and $-2\frac{1}{2}$ V respectively, the anode current will set itself at about $5\frac{3}{4}$ mA; point A on the curve.

If v_g has a peak value of 0.5 V, the total voltage on the grid will swing between −2 and −3 V, alternate half-cycles adding to or subtracting from the steady voltage E_g. The anode current will swing correspondingly with the changes in grid voltage, the points B and C marking the limits of the swing of both. The current, swinging between $7\frac{1}{2}$ and 4mA, is reduced by $1\frac{3}{4}$mA on the negative half-cycle and increased by the same amount on the positive one. The whole is therefore equivalent to the original steady current with an alternating current of $1\frac{3}{4}$mA peak superposed on it. Incidentally this, divided by the peak grid voltage, 0.5, gives 3.5 mA/V, the g_m.

There are two ways in which this a.c. in the anode circuit can be usefully employed. It can be used for producing an alternating voltage, by passing it through an impedance. If the voltage so obtained is larger than the alternating grid voltage that caused it, we have a voltage amplifier. Alternatively, if the alternating current is strong enough it can be used to operate a loudspeaker or other device, in which case the valve is described as a power amplifier. It is this ability to amplify that has made modern radio and the many kindred developments comprised in electronics possible. The subject is so important that several chapters will be devoted to it, beginning with the next one.

10.6 Grid Bias

You may have been wondering why the grid of the valve has never been shown as positive. The reason is that a positive grid attracts electrons to itself. It is, in effect, an anode. The objection to this is not so much that it robs the official anode of electrons—for up to a point it continues to increase the flow there as well as to itself—but because current flowing in the grid circuit calls for expenditure of power there, thus depriving the valve of one of its most attractive features—its ability to release or control power without demanding any from the source of input signal. For this reason valves are often called *voltage-operated* or *voltage-controlled* devices. Another way of expressing the same idea is to say that the valve, when operating normally, has an infinite input resistance. If, however, the grid is driven positive, the input resistance falls to a low value.

To prevent the flow of grid current during use, a fixed negative voltage, known as *grid bias*, is applied. The amount of bias required is equal to—or preferably a volt or so more than—the positive peak of the signal input voltage which the valve is expected to accept. The reason for making the bias greater than the signal peak is that some electrons are emitted from the cathode with sufficient energy to force a landing on the grid against a negative bias of anything up to about one volt.

10.7 The Transistor

Considering how enormously the usefulness of the thermionic valve was increased by adding a third electrode to the original diode, one is distinguished from the collector by carrying an arrow head which also distinguishes p-n-p from n-p-n types by pointing in the direction of (positive) current.

These aptly suggest the point-contact transistor (long obsolete and not mentioned in this book) but are poor representations of the bipolar transistor. However in the interests of national and international standardisation we will use these symbols throughout the book.

10.8 Transistor Characteristic Curves

To plot transistor curves we can use much the same set-up as for the triode valve (Fig. 10.1) except that there is no need for cathode heating, the supply voltages should be lower and both of the same polarity relative to the emitter, and a low-reading milliammeter is needed for base current; see Fig. 10.9. We must remember that what in Fig. 9.14 was the 'working' (forward) part of the graph has no place in connection with the collector-to-emitter circuit of a transistor, which always includes one reverse-biased junction no matter what the polarity of the voltage applied. The only place where the forward curve has any relevance is in the control (base) circuit by itself. First, though, let us deal with the collector circuit.

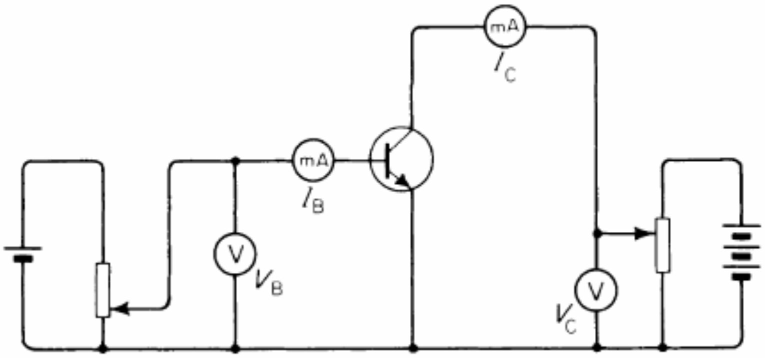


Fig. 10.9 Arrangement, similar to Fig. 10.1, for obtaining characteristic curves of a transistor

With the base disconnected, what we have is a reverse-biased p-n junction in series with a forward one. The only current that can flow through the reversed junction is the intrinsic or leakage current, which should be small enough to neglect for most practical purposes. The bias needed to make this small current flow through the forward emitter-to-base junction is also negligible. So the first curve, whether we label it '$I_B = 0$' or '$V_B = 0$', will show a small and nearly constant (over a wide range of V_C) collector current I_C.

As our circuit provides for measuring both I_B and V_B, we have a choice of two methods. We could follow the same procedure as with valves and plot I_C/V_C curves at regular intervals of positive V_B

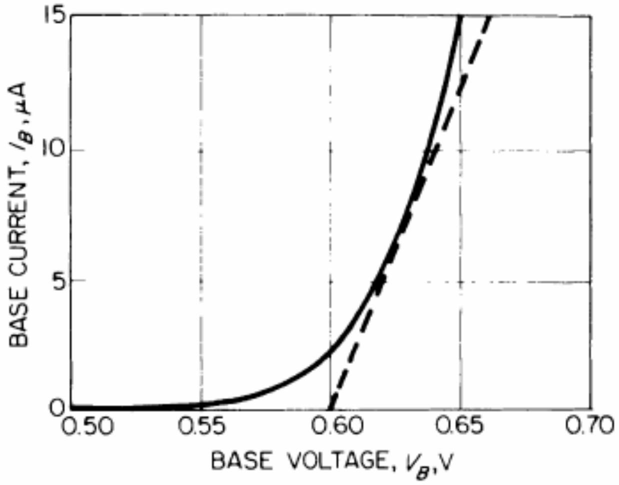


Fig. 10.13 I_B/V_B characteristic for a fixed value of V_C (0.5 V)

that it is equal to $I_C + I_B$. As both of these flow into the n-p-n transistor, I_E must flow out of it. And as I_B is normally very much less than I_C, I_E is approximately equal to I_C but a little larger.

10.9 Transistor Parameters

The most obvious differences between the I_C/V_C curves in Fig. 10.10 and the corresponding ones for a valve (Fig. 10.2) are the very rapid rise of I_C followed by a nearly flat top. In a triode valve there is a plentiful supply of electrons emitted by the cathode, and the amount of anode current depends on how much anode voltage there is to counteract the combined throttling-back effect of the negative space charge and (usually) negative grid voltage. An n-p-n transistor, on the other hand, relies on a positive base voltage to make available a supply of electrons, and a fraction of a volt at the collector is enough to get nearly all of them across the junction. Little can be done by further increasing V_C.

Although transistor characteristics differ so much from those of the triode valves seen so far, they resemble quite closely those of tetrode and pentode valves.

When considering valve curves we saw that the steeper the slope of a current/voltage characteristic the higher the conductance and the lower the resistance to current changes at that part of the curve. So Fig. 10.10*a* shows a very low resistance for the first fraction of a volt, and a very high resistance thereafter. Consider the resistance between emitter and collector terminals with 5 V_C and 620 mV V_B, so 2 mA I_C. To a.c. it can be found by noting that a change in V_C of 2.5 V brings about a change in I_C of 0.05 mA. So the a.c. resistance is $2.5/0.00005 = 50\text{k}\Omega$. To d.c., however, the resistance at the 5V/2mA point is only $5/0.002$ or 2 kΩ. The a.c. resistance, which could be denoted by r_C, is in practice seldom used, for reasons that will be explained in Chapter 12.

10.10 Field-Effect Transistors

The general class of triode transistors considered so far is sometimes called *bipolar*, because it employs charge carriers of both polarities: electrons and holes. There is another class, distinguished as *unipolar*, working on such a different principle that its right to be called a transistor at all is open to doubt, but it is nevertheless known as the field-effect transistor, usually abbreviated to f.e.t. To some extent it combines the most useful features of valves and bipolar transistors; in particular, the power economy and compactness of the transistor and the almost infinite input resistance of the valve. On the other hand it is much more easily destroyed by misuse than valves.

An f.e.t. consists essentially of a channel of semiconductor material, either p or n type, on the side of which is a control electrode of the opposite type, as shown in Fig. 10.14. Unlike the bipolar transistor, the p-n junction so formed is biased in the reverse polarity; hence its high resistance. Again, we have to learn a new set of names for the electrodes. In place of the valve cathode or transistor emitter we have the *source*. In place of the anode or collector there is the *drain*. There is however a tendency to use the terms emitter and collector for f.e.ts too. But the third electrode in a f.e.t. is always the *gate*. Increasing its reverse bias increases the depletion layer, as explained in Sec. 9.9, and thus narrows the conducting channel between source and drain, reducing the amount of current the applied voltage can pass, just like a water or gas tap. If sufficient bias is applied the channel is closed and the drain current cut off.

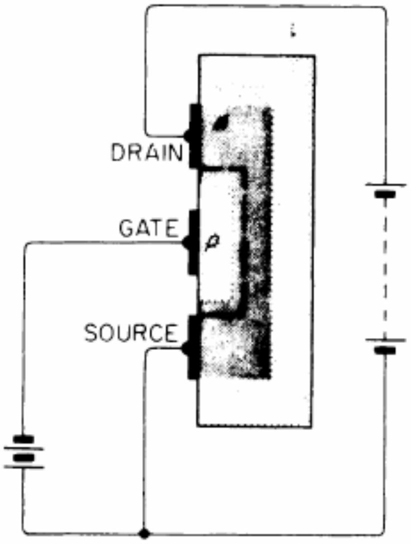


Fig. 10.14 Showing in principle how a field-effect transistor is constructed

With an n-type channel, as in Fig. 10.14, the reverse gate bias is normally negative relative to the source, and the drain voltage is positive. In both these respects, and in starting with maximum current at zero bias voltage, this f.e.t. resembles a valve. But drain-current/drain-voltage (I_D/V_D) curves, shown in Fig. 10.15, are more like transistor curves; for quite a different reason however. Increasing the drain voltage accelerates the electrons and so tends to increase the current, as with ordinary resistance, but at the same time it makes all parts of the channel progressively more positive towards the drain end, thereby increasing the reverse bias across the junction and narrowing the channel. Beyond a certain V_{D}, called the *pinch-off voltage,* the two effects almost balance, so the curve flattens out as shown. The a.c. resistance is therefore fairly high. Note however that the curves level out more gradually than in bipolar types, and in general the working drain voltages tend to be higher than collector voltages. F.e.t. mutual conductances are of the same order as in valves.; Typical I_D/V_G curves are so similar to valve I_a/V_g curves such as those in Fig. 10.3 that there is really no need to show any. In fact, f.e.ts can often be substituted for valves with little or no circuit modification.

The f.e.t. illustrated so far is an n-channel type, corresponding to an n-p-n bipolar transistor. Just as there is the alternative p-n-p type, so there is the p-channel f.e.t., requiring negative V_D and positive V_G bias. Both of them work in what is called the depletion mode, the effect of applying bias being to reduce the drain current.

10.11 Insulated-Gate F.E.Ts

There is another group of f.e.ts which offer not only a choice of polarities (n-channel or p-channel) but also a choice between the depletion mode and the enhancement mode, in which the effect of applying bias is to increase the drain current from zero or thereabouts. Instead of the resistance betwen gate and drain being in megohms as in the f.e.ts we have been considering, it is in thousands of megohms, or gigohms (GΩ), at least.

The reason for this remarkably high control resistance, greater even than that of valves, is that the gate electrode is very highly insulated. So in contrast to the junction-gate f.e.ts (jugfets) already described they are known as insulated-gate f.e.ts (igfets) or alternatively, because of a particular construction favoured, metal-oxide-silicon or metal-oxide-semiconductor f.e.ts (mosfets or mosts). This basic difference is indicated in the symbols for igfets (Fig. 10.16*b* and *c* compared with that for a jugfet *a*). All these symbols refer to n-channel types, with positive drain voltages; for p-channel the arrow heads are reversed. Both depletion and enhancement n-channel igfets are laid down on a block of p-type material (the substrate), making a fourth terminal, which may be connected straight to the source, or not, as required. Below the source and drain terminals are n-type regions, heavily doped to keep a negative bias to the gate repels electrons present in the channel, reducing its conductivity, and ultimately cutting off drain current entirely. The characteristic curves are therefore similar to those for jugfets (Fig. 10.15) except that the zero-V_G curve is usually much lower, but because the gate is insulated it can be used to increase as well as reduce the drain current. Doing this with a jugfet would drastically reduce the input resistance.

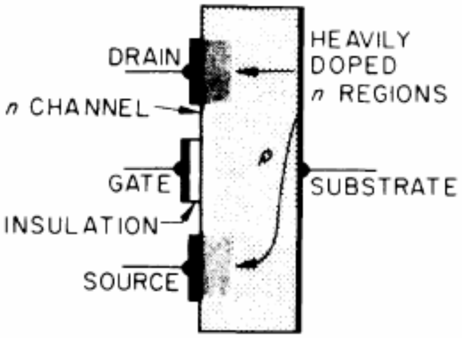


Fig. 10.17 Basic construction of n-channel depletion igfet

In an enhancement igfet the channel is missing, so with no gate bias there is no conduction between source and drain. This feature is indicated in its symbol (Fig. 10.16*c*) by the gaps between source and drain. To turn the device on, positive gate bias is needed. In this respect it resembles bipolar transistors, but in all other respects it is similar to the depletion igfet. Because the bias has to make the p-type material that lies under the gate intrinsic before it can begin to convert it into an n-type channel, several volts of bias are needed to bring the device to a suitable working point.

The gate insulation in an igfet being so extremely thin, of the order of 0.001 mm, it can be permanently broken down by quite a low voltage—25, say—and because the capacitance is no more than perhaps 4 pF the quantity of electric charge needed to destroy the device instantly is extremely small. Substituting the above figures in $Q = VC$ (Sec. 3.2), $25 \times 4 \times 10^{-12} = 10^{-10}$C. This can easily be generated by taking a step or two on a dry floor in rubber soled shoes. So precautions have to be taken, such as wrapping the device in foil until installed, and shunting it by protective diodes.

10.12 Light-Sensitive Diodes and Triodes

We have already seen that the reverse (leakage) current of a p-n diode depends very much on the temperature, because heat energy splits up atoms into free electrons and holes, both of which are charge carriers. Light has the same effect, and for this reason ordinary semiconductor diodes and transistors have opaque containers. But for many purposes small light-sensitive devices are useful. So what

CHAPTER 11

The Triode at Work

11.1 Input and Output

Although in the last chapter attention has been focused on triode valves and transistors themselves, various hints along the line (especially Sec. 10.5) must have conveyed some ideas about how these devices are employed. Now is the time to turn our whole attention to this aspect of the matter.

We have seen that basically triodes are devices by which very small amounts of electrical power—sometimes negligibly small, because requiring voltage without appreciable current—are able to control relatively large amounts, somewhat as the power of a car engine can be controlled by slight variations in pressure on a pedal. Two circuits are thus involved: the controlling or *input* circuit and the controlled or *output* circuit. Each of these must be connected to the amplifying device (such as a triode) at not less than two points. As a triode has by definition only three terminals, one of them must be common to both input and output circuits. The choice of common electrode is a very important question which we shall have to look into in the next chapter. So as not to complicate things too much right at the start, however, one particular choice has been made for each device considered: cathode, emitter and source, which in principle are all the same. This is the choice we have been assuming, because it is by far the most usual in practice, so will continue to be assumed unless one of the others is indicated.

11.2 Source and Load

Running a car engine in neutral gives it no scope for making use of the power it can develop, and is likely to be harmful if the accelerator is pressed hard down. To fulfil its purpose it needs a *load*. This term is the one used also in electronics and other branches of electrical engineering, but we must beware because it is also used to mean several things that are not quite the same—as indeed it is in general conversation (e.g., 'he's loaded'). When we say something is overloaded we usually mean that too much power is going into it and it is in danger of burning out. But here the word 'load' will be used primarily to mean whatever receives output power; for example, a loudspeaker.

Sometimes a load is fed with power direct from a generator. In this respect a television receiver as a whole is a very small part of the load connected to a generator in a power station. Simpler still, the lamp in a torch is the load fed by the battery. This book is more concerned with situations in which there is an important link between the source of power and the load; for example, a transformer, which does not itself amplify but has to be suitably adapted to its load. Or the output part of a triode, which is a link between the source of anode or collector voltage and the load. But more often the triode and any power supplies it needs are regarded as one unit (an amplifier, for example) which is a link between the controller and the load.

This controller, typified by the driver of a car, is usually called the *source* of the voltage or current needed to control the *device**, and is the essential item in its input circuit, just as the load is in the output circuit.

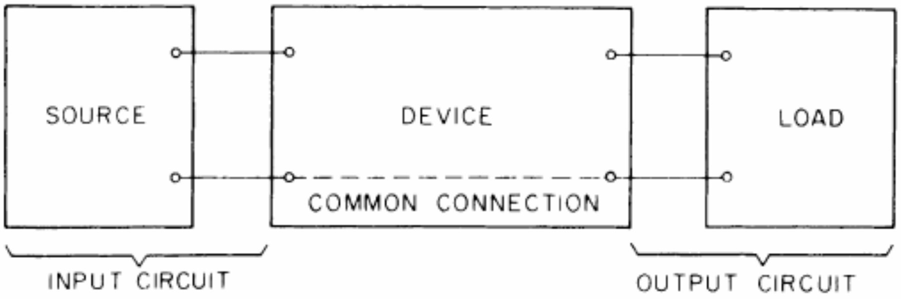


Fig. 11.1 Triodes, like many other devices and complete units for such purposes as amplification, have an input and an output, as shown here in general

The substance of these first two sections can be shown quite simply as Fig. 11.1. Later on, when we see the importance of the impedances between the four pairs of terminals, we shall have to elaborate this diagram somewhat.

11.3 Feeds and Signals

In the meantime there are a few practical details and terms that need to be made clear before we go on.

You have gathered, no doubt, that any triode (and the same is true of electronic devices with more than three electrodes) can only work properly if it is supplied with certain suitably chosen currents and voltages. What is needed can usually be ascertained with the aid of characteristic curves—together with some understanding of how the device works, which by now we are supposed to have. For instance,

---

*Components in electronics are often termed devices. Those, such as valves and transistors, which can generate signals and amplify are called *active devices*, and components which cannot (typically resistors, inductors and capacitors) are known as *passive devices*.

11.4 Load Lines

At this early stage in the study of electronic circuits we must take care not to get so involved in various incidental complications that the basic principles are not clearly seen. For instance, amplification of high (radio) frequencies is beset with considerable complications, which will be kept back until Chapter 22. So for a start we shall assume signals of low (audio) frequency. They are not only simpler to deal with but are of practical interest in radio, television, tape recorders, record players and much else.

Triode valves are now almost never used for a.f. amplification, but they are freer from sidetracking issues than transistors, so the basic principles of amplification can be demonstrated more clearly at the start using a triode valve as the active device; once grasped, these principles provide a good foundation for studying other types.

Fig. 11.2, then, shows a common-cathode triode valve in the simplest possible amplifier. It is in fact Fig. 10.4 with the addition of a 20 kΩ load resistance (R) in the anode (output) circuit. This immediately faces us with a difficulty. It is easy enough to calculate the ouptut signal current (a.c.) when R is not there; it is equal to the milliamps of anode signal current per volt of grid signal (i.e., g_m) multiplied by the grid signal voltage (v_g). In Fig. 10.4 we had $g_m = 3\frac{1}{2}$ and the peak v_g was $\frac{1}{2}$ V, so the peak I_a should be $\frac{1}{2} \times 3\frac{1}{2} = 1\frac{3}{4}$ mA, and this agrees with what we found from the I_a/V_g graph. But whenever the anode current varies in Fig. 11.2 the voltage drop across R must vary correspondingly, and as the total voltage provided in this anode circuit is fixed at 240 V, the voltage across the valve (V_a) must vary oppostiely to that across R. But the definition of g_m says V_a must be constant. So that method of calculation fails. And the alternative method, by Ohm's law, using r_a as the resistance of the valve (between anode and cathode) also breaks down, because the definition of r_a specifies constant grid voltage. What are we to do?

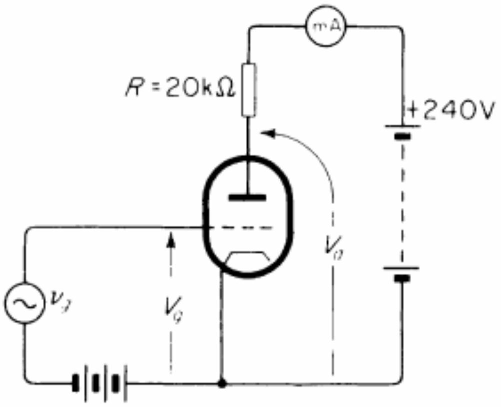


Fig. 11.2 Simple voltage-amplifier circuit

The answer is to combine the anode characteristic curves of the valve with a 'characteristic curve' of R in such a way as to make the total voltage across both of them equal to that provided, in this example 240 V. Suppose the curves in Fig. 11.3 are the ones relating

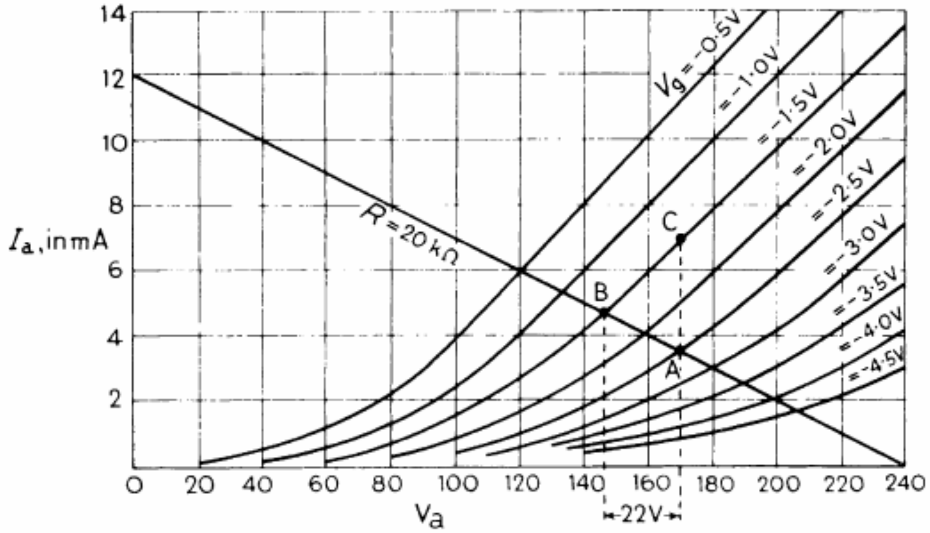


Fig. 11.3 The problem of finding the amplification of the valve in Fig. 11.2 is here worked out on an I_a/V_a curve sheet

to our valve. They tell us, of course, all the possible combinations of I_a at any V_g for which a curve is provided. When $V_g = -0.5$ V, for example, the combinations include 4 mA and 100 V, and 10 mA and 160 V, and any number of others. But in our loaded circuit the V_a corresponding to any particular I_a is fixed by the fact that V_a must be 240 V less the amount dropped in R, namely I_aR. When $I_a = 0$, then $V_a = 240 - 0, = 240$ V; when $I_a = 1$ mA, $I_aR = 20$V, so $V_a = 220$ V; when $I_a = 2$ mA, $V_a = 200$V; and so on. Plotting all these points on Fig. 11.3 we get the straight line '$R = 20$ kΩ'. Since R is a load resistance, this line is called a *load line*. It would be equally easy to draw a load line for any other value of R.

The only possible values of I_a are those that are on the appropriate load line. So now for any given load line each value of V_g has only one possible corresponding value of I_a. To check this, think of a number for I_a, say $3\frac{1}{2}$ mA. The only point corresponding to this on the 20 kΩ load line is marked A. It tells us that V_a is 170 V and V_g is -2.5 V. $3\frac{1}{2}$ mA flowing through 20 kΩ drops $3\frac{1}{2} \times 20 = 70$ V. Deducting this from the anode battery voltage, 240, leaves 170 V, which checks.

11.5 Voltage Amplification

Suppose now we alter the grid voltage to -1.5 V. The working point must move to B, because that is the only point on both the '$V_g = -1.5$ V' curve and the load line. The anode current rises from 3.5 mA to 4.6 mA: an increase of 1.1 mA. The voltage drop due to R therefore increases by $1.1 \times 20 = 22$ V. So the voltage at the anode (V_a) falls by that amount.

Note that a grid voltage change equal to 1 V has caused an anode voltage change of 22 V, so we have achieved a voltage multiplication—called amplification or *gain*—of 22. If the anode had been connected direct to a 170 V battery there could of course have been no change in V_a at all, and the working point would have moved to C, representing an I_a increase of 3.3 mA. The reason why the increase with R in circuit was only 1.1 mA was the drop of 22V in V_a, which partly offset the rise in V_g.

Another thing to note is that making the grid less negative caused the anode to become less positive. So an amplifier of this kind reverses the sign of the signal being amplified, sometimes described as inverting the signal, because if the input and output signals are depicted, one is upside down with respect to the other.

11.6 An Equivalent Generator

It would be very convenient to be able to calculate the voltage amplification when a set of curves was not available and only the valve parameters were known. To understand how this can be done, let us go back to Fig. 10.1 with its controls for varying anode and grid voltages. First, leaving the V_g control untouched, let us work the V_a control. The result, indicated on the milliammeter, is a variation in anode current. The same current variation for a given variation in V_a would be obtained if an ordinary resistance equal to r_a were substituted for the valve (Sec. 10.4). Considering only the variations (signals), and ignoring the initial V_a and I_a (feeds) needed to make the valve work over an approximately linear part of its characteristics, we can say that from the viewpoint of the anode voltage supply the valve looks like a resistance r_a.

Now keep V_a steady and vary V_g. To the surprise of the V_a supply (which does not understand valves!), I_a again starts varying. If the V_a supply could think, it would deduce that one (or both) of two things was happening: either the resistance r_a was varying, or the valve contained a source of varying e.m.f. To help it to decide between these, we could vary V_a slightly—enough to check the value of r_a by noting the resulting change in I_a—at various settings of the V_g control. Provided that we took care to keep within the most linear working condition, the value of r_a measured in this way would be at least approximately the same at all settings of V_g. So that leaves only the internal e.m.f. theory in the running. We (who do understand valves) know that varying V_g has μ times as much effect on I_a as varying a voltage directly in the anode circuit (namely, V_a).

type of diagram tends to confuse rather than elucidate. In accordance with the equation $kb = -\mu kg$, kb is drawn opposite to kg and μ times as long. It represents the generator voltage, which is divided between r_a and R. Point b is of course inaccessible in a real valve; ab is drawn dashed to emphasise this.

See Sec. 12.1 for a fuller account of equivalent generators.

11.7 Calculating Amplification

Now let us apply this equivalent generator technique to the problem we set out to solve—finding the amplification of a valve. The signal current i_a in Fig. 11.4 is (by Ohm's law) the signal e.m.f. divided by the total circuit resistance: $-\mu v_g/(R + r_a)$. The signal output voltage, v_a, is caused by this current flowing through R, so is $= \mu v_g R/(R + r_a)$. The voltage amplification (which is often denoted by A) is v_a/v_g, so

$$A = \frac{-\mu R}{R + r_a}$$

This minus sign agrees with our finding that the signal output voltage is inverted with respect to the input. Since it is generally understood that v_a is in opposite phase to v_g, the minus is often omitted. Let us apply this equation to the case we considered with the help of Fig. 11.3. Using the methods already described in Secs 10.2 and 10.4 we find that in the region of point A the valve in question has an r_a of about 14.5 kΩ and μ about 39. Substituting these, and $R = 20$ kΩ, in the formula: $A = -39 \times 20/(20 + 14.5) = -22.6$, which agrees pretty well with the figure obtained graphically. (It must be remembered that with electronic devices we are in the realm of the approximate. Performance depends so much on conditions, and even when conditions are the same, different samples of the same type of device give different results. So there is no point in specifying them with high precision.)

The amplification formula that we have just used shows at once that making $R = 0$ would make $A = 0$. It also shows that making R so large that r_a was negligible in comparison with it would make A almost equal to μ, which is the reason for calling this parameter the amplification factor. What happens at intermediate values of R can best be seen by using the formula to plot a graph of A against R—the full-line curve in Fig. 11.6, in which, incidentally, logarithmic scales (App. B. Sec. 6) have been used to cover wide ranges of values without squeezing most of them into a corner. According to this it looks as if the higher the more, so far as R and A are concerned. But although feeds do not come directly into the equivalent generator (from which this curve was derived) the assumption that they are adequate to maintain the parameters at the values assumed must never be forgotten. Looking at Fig. 11.3 we see that if R were increased to a much larger value the load line would have to swing round until it was nearly horizontal, cutting the valve curves at points where their decreased slopes would mean much larger values

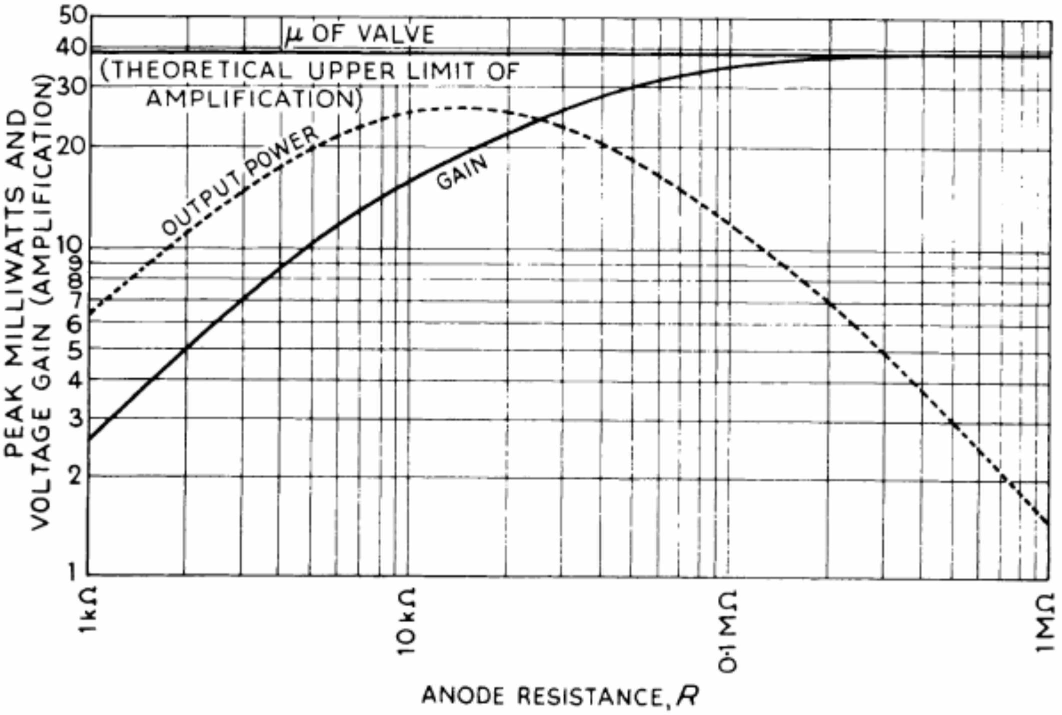


Fig. 11.6 The full line shows how the voltage amplification depends on the value of load resistance R. It is calculated for a valve with $\mu = 39$ and $r_a = 14.5\text{k}\Omega$. The dashed line shows the peak milliwatts delivered to R for a grid input of 1 V peak

of r_a too. So although very high load resistances are sometimes used, it is more likely to be for some other reason, such as keeping the feed current small, than for extracting the maximum possible voltage gain. The alternative solution, increasing the supply voltage to compensate for the loss in R, is usually uneconomic with valves, and exposes transistors to serious risk of breakdown when current is cut off. In practice there is often little or no choice of load resistance, and things may have to be designed to fit it, rather than vice versa.

11.8 The Maximum-Power Law

For some purposes (such as working loudspeakers) we are not interested in the voltage output so much as the power output. This, of course, is equal to $i_a v_a$ or $i_a^2 R$ (Sec. 2.17), and can therefore be calculated by filling in the value of i_a we found; namely $-\mu v_g/(R + r_a)$. Denoting the power by P, we therefore have

$$P = \left(\frac{\mu v_g}{R + r_a}\right)^2 R$$

A graph of this expression, for $v_g = 1$ V, and $\mu=39$ and $r_a=14.5$ kΩ

11.9 Transistor Load Lines

We have seen that the output current/voltage curves of transistors, both bipolar and field-effect, are shaped very differently from those of vacuum triodes. It happens that valves of the types that were most used, such as pentodes, also had the same sort of shape, rising very steeply at first and then flattening out. So we ought to see how the foregoing principles apply to curves of this general shape.

Fig. 11.7 is a set of output curves for a typical low-power silicon transistor. Note that this time (like Fig. 10.10*a*) they have been drawn for equal intervals of base *voltage*. As we would expect from Fig. 10.13, the corresponding steps in collector current are not at all equal. If, mindful of the maximum power law, we were to try drawing a load line having the same slope (but the other way) as the transistor lines we would see that whatever our choice the load line would have to be extended a very long way to the right before it hit the zero-current line, so that the point where it did would almost certainly represent a collector voltage above the rated maximum for the transistor. That being so, if the collector current were cut off for any reason (a large negative input signal, or a lack of bias), the transistor's life would probably come to an abrupt end. In practice one is seldom vitally interested in getting the maximum power output *per input volt or milliamp* regardless of load resistance. A commoner requirement is maximum power output for any reasonable input, and this will come up for discussion in Chapter 19.

Meanwhile, an interesting thing to notice about devices with curves of this shape is that owing to the very steep initial rise of current the load line can be extended nearly to zero V_C regardless of the slope at which it is drawn. For the same reason the maximum output voltage obtainable is practically the same for all load resistances. The three sample load lines in Fig. 11.7, based on 9 V as the collector supply voltage (V_{CC}) bring out these facts clearly. Comparing this diagram with Fig. 11.3 we see that something like twice as much of the supply voltage can be utilised.

It remains true, as with valve triodes, that the higher the load resistance the higher the voltage gain, again with the practical restriction that if we are too greedy we get into trouble with the closing up of the curves at what is called the *bottom bend*—more clearly seen as such when the relationship between I_C and V_B is expressed directly as in Fig. 10.13.

While the voltage output is limited at the lower end of the load line by bottom bend, it is stopped abruptly at the top end by coming up against the steep rise of the curves, an action that is commonly called *bottoming*. This term may appear rather an odd one for something that happens at the top end of the line, farthest from the bottom bend. The explanation is not an unseemly anatomical obsession but a reference to the output voltage, which is here at the bottom of its range—only a fraction of a volt above zero V_C. An alternative name for this state is *saturation*.

More appropriate to bipolar transistors than voltage gain is current gain. By looking at Fig. 10.10*b* and mentally drawing load lines across it we should be able to see that maximum current amplification is obtained with $R = 0$. Just as maximum voltage gain with no current is seldom useful, so is maximum current gain with no voltage. Both conditions give zero power output. However, by turning the vertical line representing zero R through a moderate anticlockwise angle we shall find that a useful voltage output is obtained without losing very much current. Try something like 500 Ω.

11.10 Class A, B and C Operation

The bias conditions so far described yield what is known as class-A amplification or operation, in which the input and output waveforms are similar, though one may be inverted. Another feature of class A is that the mean current drawn from the power supply is constant, being unaffected by changes in input-signal amplitude, even by its complete removal. Clearly such an amplifier does not use the power taken from the supply very economically!

There is, however, another mode of amplifier operation in which the active device is biased almost to the point of cut-off of output current. Only half of each input sine wave can be reproduced by such an amplifier, but, by using two similar active devices, and applying the input signal to them in such a way that one reproduces positive half-cycles and the other negative (a widely-used circuit arrangement known as push-pull), distortionless amplification can be achieved. This is known as class-B amplification. Because it is more economical of supply current it is widely used in amplifiers where that is important; e.g., in equipment run off batteries.

By further increasing the bias, the output can be made to flow only at the peaks of the input signals. In such an amplifier there is no proportionality between input and output signals. Nevertheless there are circumstances, as we shall see in later chapters, where this mode, called class C, is useful, and of course it is even more economical than class B.

CHAPTER 12

Transistor Equivalent Circuits

Don't be perturbed if you find this chapter tougher than any of the others, especially the second half, from Sec. 12.7 onward. It *is* tougher! The first half should be persevered with, and if you then feel the need for some relief you may be glad to know that a complete grasp of the second half is not essential to understanding the rest of the book. At a first approach it should be enough to study Figs. 12.14, 16, 18 and 19.

12.1 The Equivalent Current Generator

When we attempted, in Sec. 10.9, to get to grips with the leading transistor parameters on the lines that worked so well and simply with valves, we were put off by statements that what looked like the obvious equivalents for transistors were actually not very popular, and by promises that the reasons would emerge more clearly in due course. That moment has now arrived.

Two of the three main valve parameters relate output to input and are therefore called *transfer* parameters: g_m is the ratio of output signal current (i_a) to input signal voltage (v_g), output signal voltage being zero; μ is the ratio of the voltage of the imaginary output generator to the input voltage. Knowing them both, then for a given input voltage we know the maximum values of both output voltage and current. As g_m is defined for zero load resistance, the whole of that voltage is engaged in driving that current through the internal resistance of the valve, which can therefore be found as the ratio of voltage to current, or μ/g_m, and is denoted by r_a. (It can also be found directly and independently, from I_a/V_a characteristic curves.)

Both of the transfer parameters for valves refer to the input *voltage.* That is natural enough, for any input current is incidental. It is the grid voltage that does the controlling. Because the input resistance of a transistor is comparatively low, both voltage and current are needed, but our study of how a transistor worked showed that it was the base *current* that decides how much collector *current* is released. So it is the ratio of these two, the current amplification factor, h_f, that best expresses the control capability of a transistor. Like μ for a valve, h_f has the merit that its value depends less on working conditions than do most other parameters. It seems, then, that in bipolar transistors (which are the kind to be assumed unless indicated otherwise) current plays a role corresponding to voltage in terminals is whatever is needed to make that current flow. So if it is open-circuited the voltage becomes infinitely large. This kind of generator is even more difficult to imagine because nothing very like it exists. The nearest is a very high voltage in series with a very large resistance, so that inserting any reasonable load resistance makes very little difference to the amount of current. But we shall see in Chapter 27 that a constant current almost regardless of resistance (within certain limits) can be provided by special electronic circuits. Any real generator can be simulated by an ideal current generator in parallel with the appropriate resistance, or conductance.

The symbol shown in Fig. 12.3 is often used in place of 12.2*b* but is undesirable for this purpose as it is an international standard symbol for a transformer.

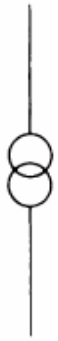


Fig. 12.3 This symbol for a current generator is deprecated

12.2 Duality

Although it will interrupt our quest for a transistor equivalent circuit, this may be as good a moment as any to take note of an important general circuit concept—*duality*. As far back as Sec. 2.12 we saw that an alternative to Ohm's law in the form $E = IR$ is $I = EG$. This is because G, the conductance, is defined as $1/R$, and substituting it in the first form of Ohm's law, after having divided both sides by R, we get the second. Comparing the two forms, we see that current and voltage have changed places, and so have resistance and conductance. This is not just a coincidence. We can extend the list:

Voltage, V or E	↔	Current, I
Resistance, R	↔	Conductance, G
Capacitance, C	↔	Inductance, L
Reactance, X	↔	Susceptance, B
Impedance, Z	↔	Admittance, Y

From any equation in which any of these quantities appear, another equation can be derived by changing each of them for the other in the

12.3 Voltage or Current Generator?

How do we decide whether to use a voltage or current generator in an equivalent circuit? Remember that although they are exactly equivalent externally they are not equivalent inside. Our domestic electricity supply could be regarded as a voltage generator of 240 V r.m.s. in series with perhaps 0.1Ω. With nothing switched on, there would be no power consumption in the equivalent generator either. The equivalent current generator would have to be 2400 A in parallel with 0.1 Ω, which would be consuming $2400^2 \times 0.1 = 576000$ W all the time!

More important for us are the terms in which the two components are expressed. For a valve the generator part can be in terms of μv_g or $g_m v_g$—voltage or current. The other part is the same in both, though in one it is normally expressed as r_a and in the other as g_a. In a triode valve particularly, r_a is usually moderate, and we could (if we used such a valve at all) make R relatively large. So for a rough approximation we could neglect r_a in comparison with R in $R + r_a$ in the equation we found in Sec. 11.7 for a valve's voltage gain:

$$A = \frac{\mu R}{R + r_a}$$

It would then reduce to the simple approximation $A \simeq \mu$. ($\simeq$ means 'is approximately equal to'.) See also Fig. 11.6. The parameter thus gives a better idea of the valve's capability than g_m, for the current generator circuit shows that when R is much greater than r_a only a small part of the generated current gets into it. When $R > r_a$, then, the voltage generator circuit is usually preferred. We have seen that both bipolar and field-effect transistors have such large a.c. output resistances that R is usually small by comparison. The same is even more true of many valves other than triodes. If then in the equation for A we neglect R in the denominator (lower part) we get

$$A \simeq \frac{\mu R}{r_a}$$

and as $\mu / r_a = g_m$ this further simplifies to

$$A \simeq g_m R$$

which is a very useful approximation. So where the load resistance is less than the device's output resistance the current generator, with g_m, is usually preferred. This preference is strengthened by the fact that in practical circuits there are a number of resistances and reactances in parallel at the output, and if they are expressed as conductances and susceptances they can be added in quadrature to find the total admittance.

12.4 Transistor Equivalent Output Circuit

Our conjecture in Sec. 12.1 that a current generator might be the one to use for representing a transistor is thus justified. The output current with $R = 0$ is, as we know, $h_f i_b$, so that is the correct output of the imaginary generator. In parallel with it is the internal resistance of the transistor, preferably expressed as a conductance. We might make its symbol g_c, but for a reason to appear later h_o has been adopted.

So now we have the two components needed for constructing the output part of a transistor equivalent circuit. But a question arises again about the sign of the generator's output. In Sec. 11.6 the reason was given for preferring a minus sign for the voltage, μv_g. It seems logical that when a current generator is substituted the sign of the current should be the same. However there is a very widely used convention that instead of a minus sign for $h_f i_b$ the direction of the current should be shown reversed, so that the positive direction of current through the load is into the collector, as in Fig. 12.4. The effect of a positive i_e is to make the emitter end of R positive, so for v_c to be positive it would have to point downwards. As the convention is to make it point upwards, for positive inputs $+ v_c$ will always have negative values. This seems confusing, but fortunately our arrowless notation does not tie one down to either convention but covers both. So Fig. 12.5*a* shows the same thing as Fig. 12.4, the convention therein being expressed by $DA = h_f i_b$. The other, $AD = -h_f i_b$, means exactly the same thing, but because *ec* is in any case equal to R multiplied by the current flowing *down* through it (*CA*) and up through the generator (*AD*) a negative value of it comes naturally as a result of the minus sign in front of $h_f i_b$, corresponding to positive values of i_b. Fig. 12.5*b* is the corresponding phasor diagram, and by comparing it with Fig. 11.5*b* we see that it is essentially the same except that voltage and current have changed places. These are therefore dual phasor diagrams. Again, a dashed line has been used for what is inaccessible in the real circuit. If you choose to regard the

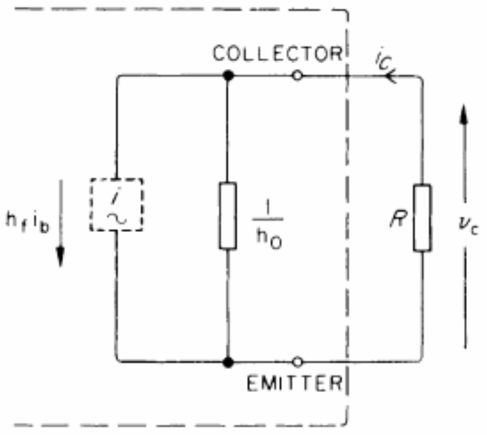


Fig. 12.4 Output part of transistor equivalent circuit using the usual conventions for *h* parameters

collector current as inwards, it is represented by *CA* (instead of *AC*), and ec shows that the voltage *ec* is negative relative to it.

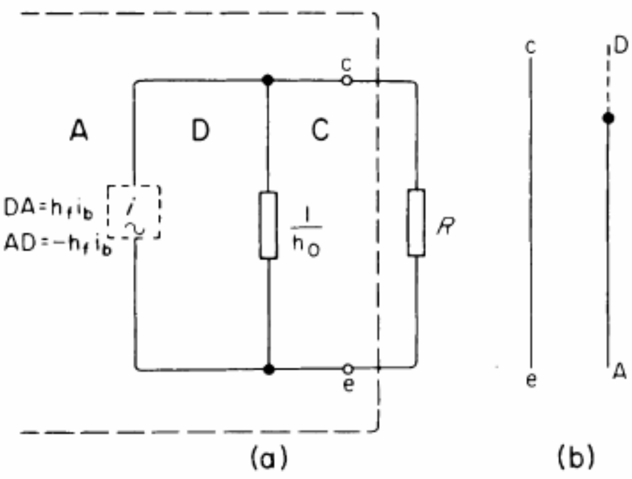


Fig. 12.5 At (*a*), Fig. 12.4 is adapted for phasor diagrams (*b*)

So far an n-p-n transistor has been assumed, but no polarity changes in equivalent circuit diagrams are needed for p-n-p types, because when they are *all* reversed they are still relatively the same. That is why it is possible to use one-way arrows for a.c. signals, which are reversing all the time. But in circuits with reactances, where phase relationships are not just simple positive or negative, arrows can be confusing, whereas the notation explained in the latter part of Chapter 5 is always unambiguous.

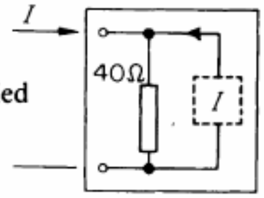
We next have to consider the input part of the transistor, which is more complicated than that of a valve, because both voltage and current are involved. The ratio of the two is the input resistance, which we found from the slope of the I_B/V_B characteristic as in Sec. 10.9. It has been given the symbol h_i, and its value depends very much on V_B, as we can tell from the variableness of the slope in Fig. 10.13. And that is not the end of the matter, for a further complication arises which involves a principle of such importance that we shall have to devote some time to it.

12.5 Some Box Tricks

Suppose we are given a 'black box' with two terminals and are told to measure the resistance between them. We do this in a simple straightforward way by applying a known voltage and measuring the current going through the box. Suppose we try 6 V and the current meter reads 50 mA. We conclude that the box contains a resistance of $6/0.05 = 120\ \Omega$. However, just to make sure we try 12 V, and expecting the current to be 100 mA we are disconcerted to find it is 200 mA. After having just been thinking of a non-linear characteristic curve, we may put forward the theory that the box an 80 Ω resistor. There would be no means, without opening the box or X-raying it, of telling whether it contained just an 80 Ω resistor or some other resistance and a source of e.m.f. proportional to that applied. And we can easily see that if the internal source in Fig. 12.8 was reversed, so that it increased the flow of current, it would make the resistance appear less than it really was.

Or imagine again that instead of being in series with a proportional voltage source our resistor was in parallel with a proportional current source, as in Fig. 12.9. For every milliamp we put into the resistor from outside, the unseen current generator puts another through it in the same direction. So the voltage drop across the resistor, which we measure between the terminals, is twice what it should be with 40 Ω. It would be reasonable to deduce, therefore, that the box contained an 80 Ω resistor. If the internal source sent current through the resistor in the reverse direction it would make the resistance seem less than it actually was.

Fig. 12.9 Box that includes a current source equal to that applied



Here we have another example of duality: voltage in opposition simulates higher resistance; current in opposition simulates higher conductance. And conversely for assisting direction.

Let us call the apparent resistance, as inferred from the externally applied voltage V and current I, R_{app}, and the actual resistance in the box, R_{act}. Let v be the internal voltage aiding V (negative value if it is opposing it). Then

$$(1) \quad R_{\text{app}} = \frac{V}{I} \quad \text{and} \quad (2) \quad R_{\text{act}} = \frac{V + v}{I}$$

From (2), $I = \dfrac{V + v}{R_{\text{act}}}$, and substituting this in (1) we get

$$R_{\text{app}} = R_{\text{act}} \frac{V}{V + v}, \quad \text{or} \quad \frac{R_{\text{act}}}{1 + v/V}$$

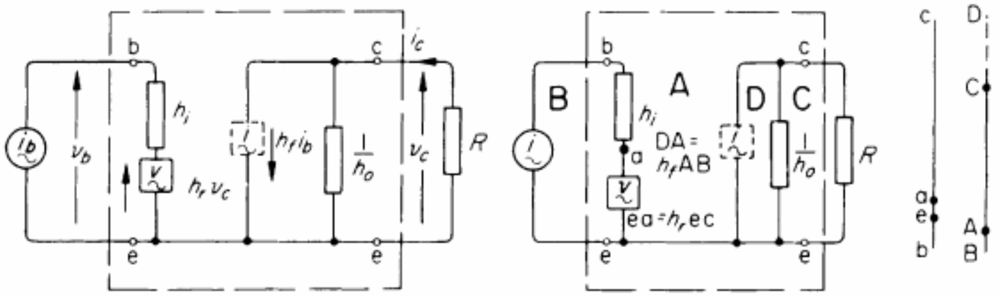
Check this with our example. Then use duality and get

$$G_{\text{app}} = \frac{G_{\text{act}}}{1 + i/I}, \quad \text{or} \quad R_{\text{app}} = R_{\text{act}} (1 + i/I)$$

where i is an unseen current augmenting I, in Fig. 12.9 for example.

12.6 Input Resistance

The previous section with its imaginary voltage and current sources adjusting themselves as if by magic to fixed proportions of whatever ea is in the same direction as ec; and ab, representing the voltage applied to h_i to cause *AB,* is in phase with it.



(a) (b) (c)

Fig. 12.12 The complete *h*-parameter transistor equivalent circuit is shown at (*a*) in conventional form and at (*b*) adapted for phasor diagrams (*c*)

12.7 Complete Transistor Equivalent Circuit

We now have a complete equivalent circuit; one of many, but worthy of a special place because its four *h* parameters are measurable practically. By the way, the *h* stands for hybrid, because they are a mixed bag: h_i is a resistance, h_o a conductance, h_f a current ratio and h_r a voltage ratio.

We have seen that when measuring h_i the load resistance must be kept at zero, or as near it as makes little difference. This obviously applies also to measuring h_f so that all the current passes through the external measuring instrument. When measuring h_r, by applying a signal voltage to the output terminals and observing the resulting voltage (or more likely microvoltage) at the input terminals, no signal current must be allowed to flow in that circuit, or some of the voltage would be lost in r_b. Lastly, h_o can also be measured by applying a voltage to the output terminals and noting the current that goes in. We would again have to make sure that no feedback signal current flows in the input circuit, because an amplified current due for it would pass through h_o and confuse the measurement. (For the same reason, when plotting I_C/V_C curves for fixed values of I_B it may be necessary to adjust V_B slightly to keep I_B constant.)

Just as the input resistance is exactly equal to h_i only when there is zero output voltage and therefore nothing from the internal voltage generator, so the output conductance is exactly equal to h_o only when there is zero input current and therefore nothing from the internal current generator. (Note duality again.) And just as the effective input resistance is in general not h_i but depends on the load resistance (for that is what gives rise to the v_c or *ea* factor in the reverse feedback voltage), so the effective output conductance is in general not h_o but depends on the signal source conductance (for that is what gives rise to the i_b or *AD* factor in the forward transfer current).

Suppose we try to measure h_o and the condition of zero input source conductance is not fulfilled. Then the measuring current applied to the output, say by making *ec* positive so that it flows into c, develops a positive $h_1 ec$ which drives a positive current *BA* round the input circuit. This is a negative *AB,* so *AD* is negative twice and therefore positive. Some of this current goes through h_o, in the same direction as that supplied. So (see end of Sec. 12.5) it makes the output conductance appear less than h_o. *So the greater the load resistance the less the input resistance, and the greater the input source conductance the less the output conductance.*

12.8 A Simpler Equivalent

It is because bipolar transistor input and output circuits react on one another in this way that calculating their circuit performance is more complicated than for valves and f.e.ts. But if the load resistance, which we should now distinguish as R_L, is considerably less than $1/h_o$, as it very often is, and the input signal source or generator resistance (R_G) is considerably greater than h_i, as it often is or can be made, then the effect of h_r is small and its voltage generator can be omitted, as in Fig. 12.13, and calculations are quite easy. Even if this simplification gave rise to errors of the order of 30% these could be acceptable if we kept in mind that the parameters of individual transistors of the same type are likely to differ even more widely. We shall see later (Sec. 13.1, for example) that one of the objects of good circuit design is to minimise the influence of transistor parameters on the results.

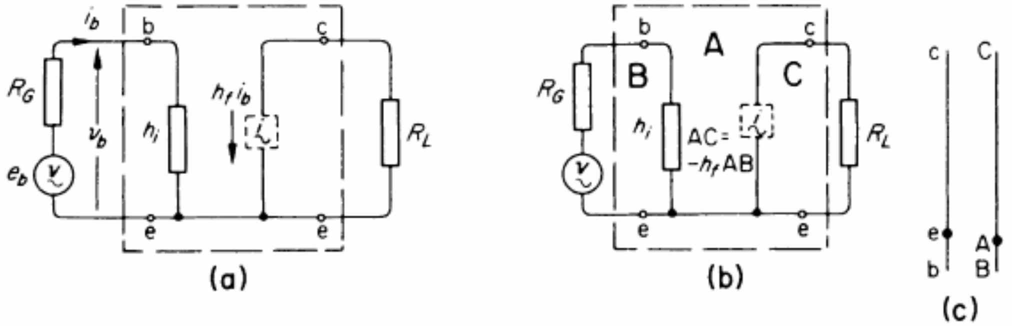


Fig. 12.13 Simplified transistor equivalent circuit, good enough in many cases. The three diagrams correspond with those in Fig. 12.12

To get the feel of the thing let us try a few figures. Suppose h_i is $3\,\mathrm{k}\Omega$, h_f is 80 and R_L is $4\,\mathrm{k}\Omega$. The approximate current gain, A_i, is simply h_f, so is 80. The voltage gain is

$$A_v = \frac{v_c}{v_b} \simeq -\frac{i_c R_L}{i_b h_i} = -\frac{h_f R_L}{h_i} = -\frac{80 \times 4}{3} = -107$$

The input resistance, r_{in}, $\simeq h_i$, and the output conductance, g_{out}, would on our assumption be zero, but if h_o is known it might as well be used as an approximate g_{out}. And we get better approximations for A_i and A_v, without adding much to the work, by taking account of it. Suppose it is 25 µS ($= 25 \times 10^{-6}$ S). The load conductance, G_L, $= 1/R_L$, is 1/4000, or 250 µS. So the total conductance into which the output current goes is 275 µS, and the load's share being 250 our figure for A_i is reduced to $80 \times 250/275$, or 73. A_v is reduced in proportion, to 97.

Compare these approximations with the results using the full equivalent circuit. At this stage the derivation of the full formulae would be rather too much, so here they are just stated:

$$A_i = \frac{h_f}{h_o R_L + 1} \qquad A_v = \frac{-h_f}{h_i(h_o + G_L) - h_r h_f}$$

$$r_{in} = h_i - \frac{h_r h_f}{h_o + G_L} \qquad g_{out} = h_o - \frac{h_r h_f}{h_i + R_G}$$

Suppose h_r is 5×10^{-4} and R_G is 5 kΩ. The results of filling in the values in the above, with the previous approximate results for comparison, are

	Without h_r or h_o	*Without h_r*	*Full working*
A_i	80	73	73
A_v	−107	−97	−102
r_{in}	3 kΩ	3 kΩ	2.855 kΩ
g_{out}	0	25 µS	20 µS

So the simplest diagram gives very good approximation (for this kind of work) to A_v, and with h_o gives correct A_i. r_{in} is not much less than h_i; but g_{out} is rather more affected by the finite value of R_G, though still not enough to worry about. Even without allowing for h_o the answers, obtainable in a few seconds, are good enough for a preliminary estimate.

Sometimes the values of r_b and r_e (Fig. 12.10) are given instead of h_i and h_r. The working r_{in} is then $r_b + (h_f + 1)r_e$ as explained in Sec. 12.6, and as 1 is usually negligible compared with h_f the fomula can be simplified to $r_b + h_f r_e$. Given h_f (and perhaps h_o) the current and voltage gains are calculable as before.

12.9 Other Circuit Configurations: Common Collector

So far we have considered only the circuit configuration or mode in which the emitter (or cathode, or source) is the electrode common to both input and output. As a triode has two other electrodes there are two other possibilities—unless we make three more by interchanging

The reason for A_v being always less than 1 is very clear in the phasor diagram. We could draw phasor diagrams corresponding to Figs 12.12 and 12.15*b* but they will be clearer if they are referred to the actual transistor circuits, as in Fig. 12.16. (Helpful though equivalent circuits are as a logical basis for calculations, we must never become so at-home in their world of make-believe that we forget that they are not real circuits, any more than the algebraical symbols.) The phasor diagrams for common-emitter and common-collector turn out to be identical. The only distinction we can make is to show which is the common terminal by encircling its point. We must remember that so far as the transistor is concerned the working input terminals are always e and b. In Fig. 12.16*a* these also happen to be the external input terminals, but in Fig. 12.16*b* the input terminals embrace the output voltage too, so the gross input must always be as much more than the output as is needed to provide the working input. This is a result of 100% voltage feedback. The voltage phasors in Fig. 12.16 show that whereas in *a* the input and output are in opposite directions and very different in magnitude, in *b* the emitter output potential is nearly the same in both magnitude and direction as the input, so a transistor in this mode is often called an *emitter follower.* The emitter (output) potential follows that of the base (input). Similarly the common-anode valve is called a cathode follower.

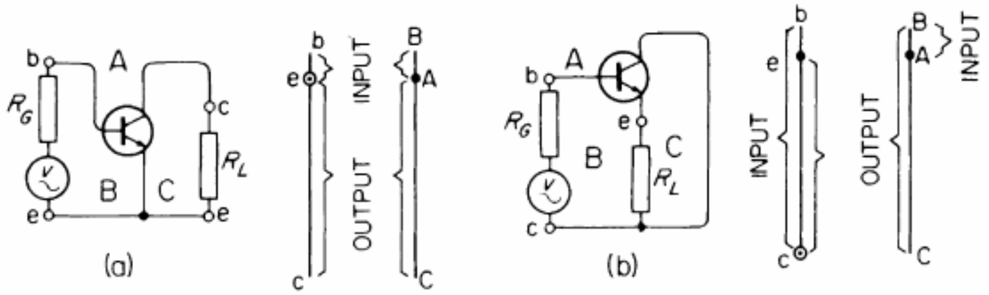


Fig. 12.16 Basic transistor amplifier circuits and their phasor diagrams: (*a*) common-emitter: (*b*) common-collector. Note that the phasor diagrams themselves are identical

Substituting various values of R_L and R_G in the formulae shows that the input and output resistances depend very much on them. When we have dealt with the common-base mode we shall look at some comparative curves for all three modes.

12.10 Common Base

Although now perhaps the least used of the three modes, historically this was the first. Whereas the common collector has 100% voltage feedback, common base has 100% current feedback. The that is to say, the ratio of load resistance to zero-R_L input resistance. And if R_G is large, g_{out} is not much more than $h_{ob,}$, but in practice there could be an appreciable difference. Note the much lower r_{in} and g_{out} than with common emitter. The contrast with common-collector is even greater. So choice of mode (and to some extent of R_G and R_L) gives one a wide range of control over input and output resistances of a transistor.

We do not really need another diagram to compare with Figs. 12.12*a* and 12.15*b*, for the only differences would be the terminal labels and the second subscripts to the *h* parameters. But Fig. 12.18 is for comparison with Fig. 12.16, and again we see the phasor diagram is the same except for the encircled point marking the common electrode. The voltage diagram shows at once that only common emitter gives a signal inversion. Study of the three phasor diagrams helps one to see how the common-emitter gives both voltage and current gain, whereas common-collector gives only current gain and common-base only voltage gain. The obtainable power gain of the common-emitter mode is therefore obviously much greater than that of the other two, which accounts for its popularity. But it must be re-emphasized that we are assuming low-frequency signals.

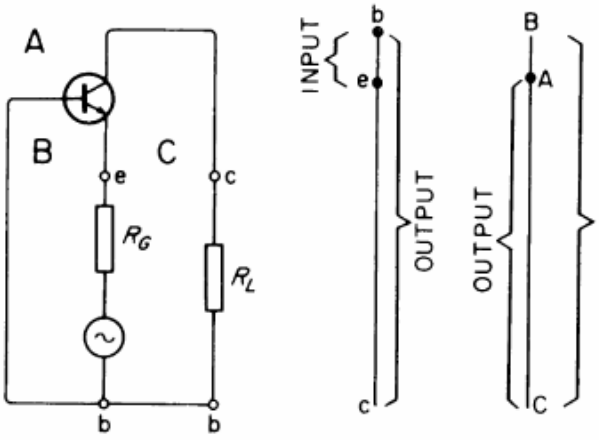


Fig. 12.18 Common-base amplifier circuit and phasor diagrams for comparison with those for the other modes (Fig. 12.16)

Note that A_i and A_v are in terms of the input current and voltage actually reaching the input terminals, not those supplied by the input current or voltage signal generator. So although in our numerical example the common-base mode had the same A_v as the common-emitter, the higher input current with common-base would mean far greater loss in the generator resistance, R_G.

12.11 A Summary of Results

Perhaps the best way of taking in the results of the last three sections is to look at Fig. 12.19. This shows how current, voltage and power

$$\text{Current gain, } A_\text{i} = \frac{h_\text{f}}{h_\text{o}R_\text{L} + 1}$$

$$\text{Voltage gain, } A_\text{v} = \frac{-h_\text{f}}{h_\text{i}(h_\text{o} + G_\text{L}) - h_\text{r}h_\text{f}}$$

$$\text{Power gain, } A_\text{p} = A_\text{i}A_\text{v}$$

$$\text{Input resistance, } r_\text{in} = h_\text{i} - \frac{h_\text{r}h_\text{f}}{h_\text{o} + G_\text{L}}$$

$$\text{Output conductance, } g_\text{out} = h_\text{o} - \frac{h_\text{r}h_\text{f}}{h_\text{i} + R_\text{G}}$$

$$\text{Output resistance, } r_\text{out} = \frac{1}{g_\text{out}}$$

APPROXIMATIONS (R_L or $R_\text{G} = 10^2 - 10^4$ unless stated)

	common emitter	*common collector*	*common base*
A_i	h_fe	h_fc	1 (R_L up to 10^5)
A_v	$h_\text{fe}\,\dfrac{R_\text{L}}{h_\text{ie}}$	1 (R_L above 10)	$h_\text{fe}\dfrac{R_\text{L}}{h_\text{ie}}$
A_p	$h_\text{fe}^2\,\dfrac{R_\text{L}}{h_\text{ie}}$	A_i (R_L $10^3 - 10^4$)	A_v (R_L up to 10^5)
r_in	h_ie	$-h_\text{rc}h_\text{fc}R_\text{L}$	h_ib
g_out	h_o (R_G above 3×10^3)	$-\dfrac{h_\text{rc}h_\text{fc}}{h_\text{ic} + R_\text{G}}$	h_ob (R_L above 3×10^3)

12.12 F.E.Ts and Valves

With their near-infinite low-frequency input resistance, field-effect transistors resemble valves much more than they resemble the bipolar transistors we have been considering, so that the same simple equivalent circuits and gain calculations studied for valves in Chapter 11 will do for f.e.ts. If the valves are of the pentode type the output-current/output-voltage curves are very similar too and in this respect they both differ little from bipolar transistors except for the greater voltage before the nearly flat (high-resistance) parts are

CHAPTER 13

The Working Point

13.1 Feed Requirements

We have been studying at some length the parameters and performance of transistors and valves, but of course these qualities cannot be obtained unless the devices are fed with d.c. power adjusted to achieve a suitable working point. What is suitable depends on the kind of work they will be required to do, and for a start we shall continue to assume low-frequency low-power amplification. Modifications for other kinds of work will appear when we come to them.

First of all let us deal with n-p-n transistors. The same information can be applied to p-n-p types if all polarities are reversed. F.e.ts and valves will follow.

As we have seen, the emitter-to-base junction is a diode which under working conditions is forward biased. The easily remembered rule for this is that *p*ositive bias supply goes to the *p* side of the junction. The collector-to-base junction is always reverse biased. (A transistor as a whole is more than just two diodes connected together, of course.) With reference to the emitter, the collector and base biases therefore have the same polarity. In this respect they differ from valves and most f.e.ts, in which the control diode is reverse biased (negative to gate or grid). The valve diode, which has the cathode as one of its electrodes, is forward biased, with positive to anode. (With valves at least, the term 'bias' is usually restricted to the grid supply.)

Fig. 13.1 summarises these requirements. It also shows the symbols used for feed voltages (and currents, where applicable): a capital V (or I) with double capital subscripts to indicate the electrode for which intended, followed if necessary by a third letter to indicate the reference electrode. (Note, this is opposite to the convention used for phasor diagrams.) K rather than C is used for cathode, as C is reserved for collector or capacitance. As anything to do with signals is omitted, there is no need to show common-collector, etc., arrangements separately; so far as feeds are concerned they are the same as common-emitter, though the actual voltages required may differ. For the same reason the feed voltages look the same as those actually applied to the electrodes, but where loads, etc., intervene, the electrode voltages are distinguished by single subscripts as used in previous chapters.

Although two batteries are shown for each device, arrangements

The answer to this problem is so to design the circuit that any changes tending to displace the working point are automatically counteracted to a sufficient extent to maintain performance reasonably close to that intended. To do this we must know about the causes of working point displacement.

13.2 Effect of Amplification Factor Variations

Fig. 13.2 shows a typical I_C/V_C curve sheet with load line. P has been selected as the working point because it is in the middle of the usable part of the load line, so that positive and negative input half-cycles, each of 20 μA peak value, could be accepted and would yield apparently equal output half-cycles, each of about 5½ V peak.

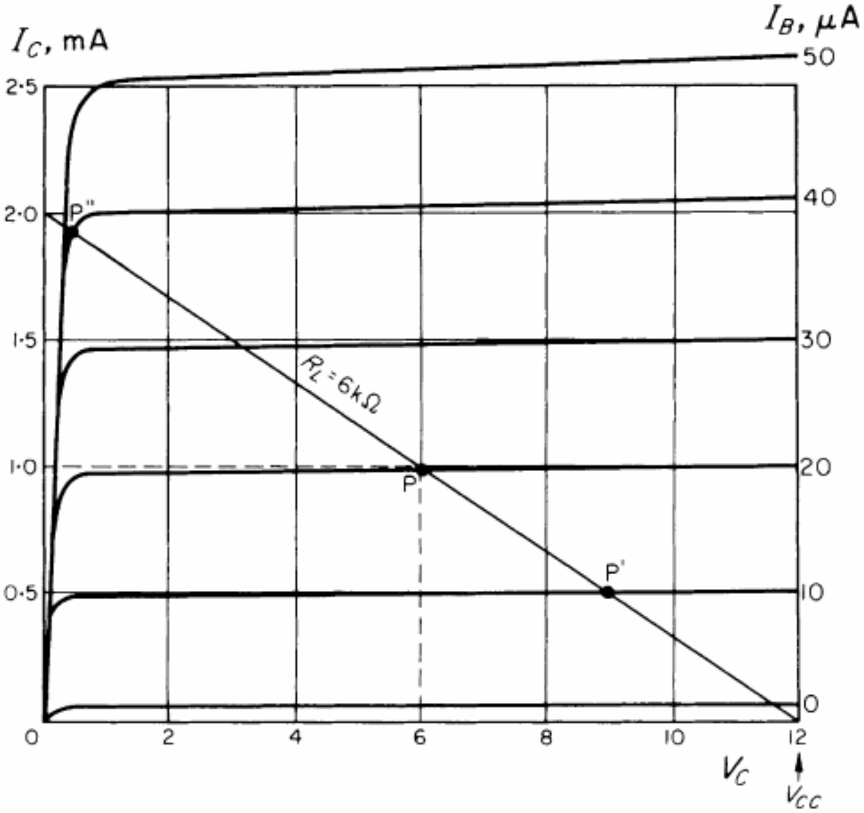


Fig. 13.2 Typical I_C/V_C curves and load line for low-power transistor. The designed working point (P) may be drastically altered (to P′ or P″, for example) by changes in h_{fe}

The curves tell us that for each increase of 10μA in I_B the collector current increases by about 500 μA, so h_{fe} is about 50. If this was classed as 'typical', for others of the same type it could range from 25 to 100. With a 25h_{fe} the 20μA I_B line would be about where the 10 μA line is in Fig. 13.2, so the working point would be displaced to P′, halving the maximum signal that could be handled without cutting off its peaks. With a 100 h_{fe} sample the 20μA line would be about where the 40μA line is in Fig. 13.2, so the working point would be

P″ and while negative half-cycles could be handled splendidly the positive halves would be cut off almost completely, since V_C is only a fraction of a volt. It looks as if our aim should be to maintain the designed working collector current rather than base current.

13.3 Influence of Leakage Current

Although leakage current came into Chapter 10 it has been ignored since, because it has no place in equivalent circuits that are concerned only with signals. Even now it may not appear to be important, for it is represented in graphs like Fig. 13.2 by the height of the zero-I_B line above the zero-I_C line, and unless it is exaggerated it is often too small to show clearly, even for a germanium transistor, let alone silicon. All it seems to do is shorten to an insignificant extent the useful part of the load line at the high-V_C end. But that is leaving temperature out of account. And we must remember that the temperature concerned is that of the junction itself, which is affected not only by the external ('ambient') conditions but by the electrical power dissipated in the transistor. Temperature is what limits the power a transistor can handle, and the reason is the rapid increase in leakage current that it causes.

There are actually three leakage currents, depending on which of the three electrodes is open-circuited, or at least reduced to zero current. Of these, I_{EBO}, the current from emitter to base with collector unconnected, is unimportant, because this diode is normally forward biased. I_{CBO} is the reverse current of the collector-base diode, as shown in Fig. 13.3*a*, and is the one most often given on data sheets. The leakage current represented by the $I_B = 0$ line in Fig. 13.2 is I_{CEO}, which passes through both diodes (Fig. 13.3*b*) and is the most important in practice. Note that the reverse voltage used for measuring leakage current is not specified. Provided it is not less than say 1 V and is not high enough to threaten breakdown, its value hardly matters, as Fig. 13.2 shows.

Thoughtful readers looking at Fig. 13.3*b* may be puzzled. It doesn't seem to conform to the rule that $I_C = h_{FE} I_B$, for here I_B is zero but I_C is not. Observant readers will have noticed that although the voltage between collector and base in Fig. 13.3*a* has the normal working polarity for an n-p-n transistor the base current is flowing the wrong way for normal operation. If an adjustable bias were added between base and emitter, as at *c*, there would be an adjustment at which the currents into and out from the base were equal, leaving zero net current and so reproducing the conditions of diagram *b*. The current going into the base, which we have made equal and opposite to I_{CBO}, by transistor action causes h_{FE} times as much current to flow to the collector. The total collector current, which in these circumstances is by definition I_{CEO}, is thus equal to

$I_{CBO}(h_{FE} + 1)$. It is therefore many times larger than I_{CBO}. Note the capital letter subscripts in h_{FE}, signifying that it is the d.c. no-signal ratio I_C/I_B, not the signal ratio i_c/i_b as in the last chapter. h_{FE} and h_{fe} are usually of the same order, at least; but both depend somewhat on the working point and temperature.

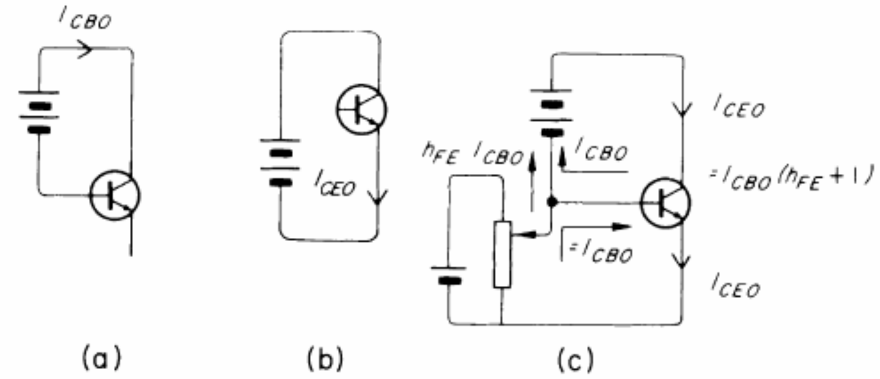


Fig. 13.3 For the elucidation of leakage currents in transistors

A typical value of I_{CBO} for a low-power germanium transistor at 25° C is 3 μA. I_{CEO} could therefore be something like 200 μA, or 0.2 mA. As the leakage doubles for about every 9° C, at 70° C this would have risen to almost 6 mA. In a case such as that shown in Fig. 13.2, lifting the whole array of I_B lines so that the zero one was at about 6 mA I_C would result in the working point being buried in the nearly vertical saturation part of the curves, putting the transistor virtually out of action. Negative base bias would be needed.

Again, it seems clear that stabilising collector current, rather than base current, should be the aim.

13.4 Methods of Base Biasing

The simplest method of providing base bias current is from the collector supply through a resistor, R_B in Fig. 13.4. Suppose for example that the required bias is 30 μA and V_{CCE} is 9 V; then R_B should be $9/30 = 0.3\,\mathrm{M\Omega}$. In this calculation the difference between the potential of the base and that of the emitter has been neglected. As we noted in Sec. 9.12, the forward voltage bias for a germanium diode is about 0.2 V and for silicon 0.6 V, and because the current rises steeply (see Fig. 10.13, for example), which means that the g_m of bipolar transistors is high, the working range of bias voltage is quite small and never far from the above values. For d.v. calculations the base-to-emitter voltage can therefore be assumed to be approximately these values. Anyone who wanted to calculate R_B more accurately by deducting the appropriate p.d. from V_{CCE} could do so, but he would be wasting his time because nothing about transistors is as precise as that. In any case in practice the choice of current-supplying source and is not always convenient. The only other one that perhaps deserves showing now is outlined in Fig. 13.7. The object of connecting R_B to the collector instead of as in Fig. 13.4 is to compensate in a simpler way for changes in I_C. If I_C is too large, it increases the voltage drop in R_L and so reduces the voltage applied to R_B, reducing I_B and therefore I_C; and oppositely if I_C is too small. There are several disadvantages, however, one of them being that signal as well as feed current is fed back from collector to base. For these reasons the circuit is not used in this simple form, but in more sophisticated circuits using pairs of transistors (described in a later chapter) these drawbacks are avoided. So successful are these circuits that i.cs using them can maintain a satisfactory performance at temperatures as low as −15°C and as high as +70°C.

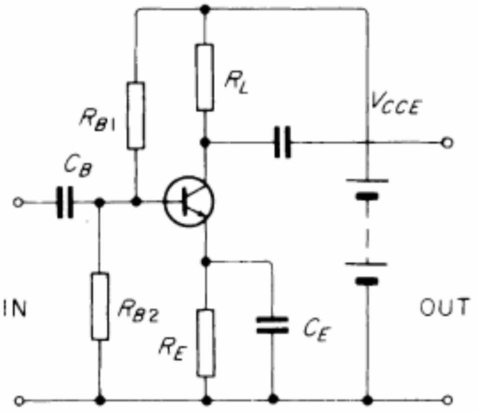


Fig. 13.6 This base-biasing scheme is more stable than the one shown in Fig. 13.4 and is usually preferred

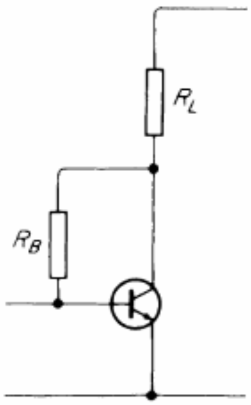


Fig. 13.7 This base-biasing circuit is much simpler than Fig. 13.6 and does have stabilising action, but owning to certain disadvantages is less used

Common-collector transistors call for no separate attention as they automatically provide compensated bias of the Fig. 13.6 type to a maximum degree.

There is no difference in principle in biasing transistors in the common-base mode; for example, Fig. 13.6 can be converted by short-circuiting the input terminals (so that C_B earths the base to signals), the input signal source being connected between emitter and R_E.

13.5 Biasing the F.E.T.

The shape of the f.e.t. output characteristic curves (Fig. 10.15, for example) is similar to those for bipolar transistors, and so is the drawing of load lines and finding a working point, allowance being made for the higher saturation or bottoming voltage; about 5V in the example referred to. (Drain voltages are usually higher than collector voltages, too.) But when biasing a depletion-mode device normally have to be several times higher in resistance than R_G, and such values are not usually very satisfactory in situations such as this. To enable conveniently lower values to be used without losing the potentially high input resistance of the f.e.t., a single high resistance can be connected between the junction of the potential divider and the gate.

13.6 Valve Biasing

Hardly anything need be said under this heading at the present stage, because the methods for f.e.ts can be used for valves, the simpler form (Fig. 13.8) usually giving adequate compensation. Other methods for both valves and f.e.ts are commonly used however, but they will be easier to understand in connection with the applications considered later (Sec. 27.8).

13.7 Biasing by Diodes

In many circuits there is a need for a fairly small constant bias voltage. Provided one can make do with about 0.2 V or 0.65 V or a combination of these, a convenient, reliable and cheap method is to pass a current through a germanium or silicon junction diode or through two or more in series, in the forward or low-resistance direction, and use the voltage drop across it or them. As Fig. 13.9 shows, the precise current used does not matter, nor does it have to be anywhere near constant. Fig. 19.15 is an example of this technique, but see Sec. 27.9 for the production of higher constant voltages by special reverse-connected diodes. Fig. 13.9 is, of course, the diode forward characteristic (e.g. Fig. 9.14) turned through a right angle.

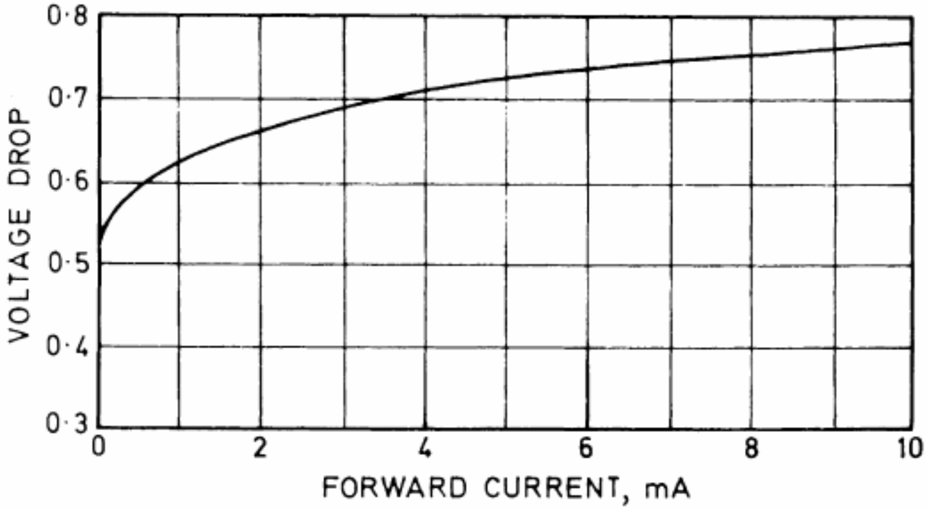


Fig. 13.9 Showing how the voltage across a typical silicon junction diode changes little with the current through it, over a wide range

CHAPTER 14

Oscillation

14.1 Generators

Having now a foundational knowledge of the few basic electrical and electronic circuit elements, we can begin to look at their applications. Elementary amplifiers have already appeared, and frequent references have been made to signals. Generators of signals have so far been represented by conventional symbols. The only sort of a.c. generator actually shown (Fig. 6.2) is incapable of working at the very high frequencies needed in radio and other systems, and produces a square waveform which for many purposes is less desirable than the sine shape (Fig. 5.2).

The alternating currents of power frequency—usually 50 Hz in Great Britain and many other countries—are generated by rotating machinery which moves conductors through magnetic fields produced by electromagnets (Sec. 4.3). Such a method is quite impracticable for frequencies of millions per second. This is where the curious properties of inductance and capacitance come to the rescue.

14.2 The Oscillatory Circuit

Fig. 14.1 shows a simple circuit comprising an inductor and a capacitor in parallel, connected as the load of a f.e.t. The f.e.t. is not essential in the first stage of the experiment, which could be performed with a battery and switch, but it will be needed later on. A valve would do just as well, or (with minor modifications) a bipolar transistor. With the switch as shown, there is no negative bias, and we can assume that the current flowing through the coil L is fairly

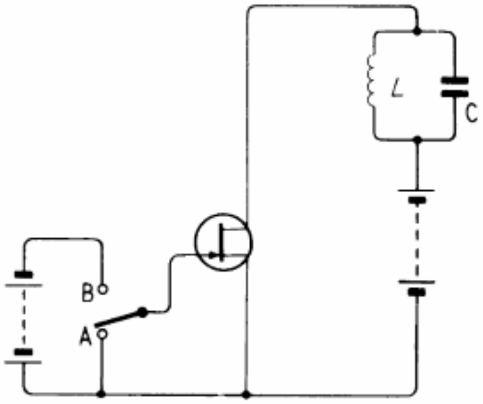


Fig. 14.1 Apparatus for setting up oscillations in the circuit *LC*

current is zero there can no longer be any magnetic field; its energy has been transferred to C as an electric field.

The voltage across C must be balanced by an equal voltage across L, which can result only from the current through L continuing to fall; that means it must become negative, reversing its direction. The capacitor must therefore be discharging through L. Directly it begins to do so it inevitably loses volts, just as punctured tyre loses pressure. So the rate of change of current (which causes the voltage across L) gradually slackens until, when C is completely discharged and there is zero voltage, the current is momentarily steady at a large negative value (point *c*). The whole of the energy stored in the electric field between the plates of C has been returned to L, but in the opposite polarity. Assuming no loss in the to-and-fro movement of current (such as would be caused by resistance in the circuit), it has now reached a value equal and opposite to the positive value with which it started.

This current now starts to charge C, and the whole process just described is repeated (but in the opposite polarity) through *d* to *e*. This second reversal of polarity has brought us to exactly the same condition as at the starting point *a*. And so the whole thing begins all over again and continues to do so indefinitely, the original store of energy drawn from the battery being forever exchanged between coil and capacitor; and the current will never cease oscillating. Note that from the moment the switch was moved the f.e.t. played no part whatever. The combination of L and C is called an *oscillatory circuit.*

It can be shown mathematically that voltage and current vary sinusoidally.

14.3 Frequency of Oscillation

From the moment the f.e.t. current was switched off in Fig. 14.1 the current in L was bound to be always equal to the current in C because, of course, it was the same current flowing through both. Obviously, too, the voltage across them must always be equal. For a sinusoidal current, reactance is the ratio of voltage to current (Secs. 6.2, 7.2), and it follows that the rectances of L and C must be equal, that is to say $2\pi fL = 1/2\pi fC$. Rearranging this we get

$$f_0 = 1/2\pi\surd(LC)$$

where f_0 denotes frequency of oscillation. This is the same as the formula we have already had (Sec. 8.7) for the frequency of resonance. It means that if energy is once imparted to an oscillatory circuit (which is now seen to be the same thing as a tuned circuit) it goes on oscillating at a frequency called the natural frequency of oscillation, which depends entirely on the values of inductance and capacitance. By making these very small—using only a few small turns of wire and small widely-spaced plates—very high frequencies can be generated, running if necessary into hundreds of millions per second, or even more with combined structures (Secs. 8.17, 16.10).

14.4 Damping

The non-stop oscillator just described is as imaginary as perpetual motion. It is impossible to construct an oscillatory circuit entirely without resistance of any kind. Even if it were possible it would be merely a curiosity, without much practical value as a generator, because it would be unable to yield more than the limited amount of energy with which it was originally furnished. After that was exhausted it would stop.

The inevitable resistance of all circuits—using the term resistance in its widest sense (Sec. 2.18)—dissipates a proportion of the original energy during each cycle, so each is less than its predecessor, as in Fig. 14.3. This effect of resistance is called *damping*; a highly damped oscillatory circuit is one in which the oscillations die out quickly.

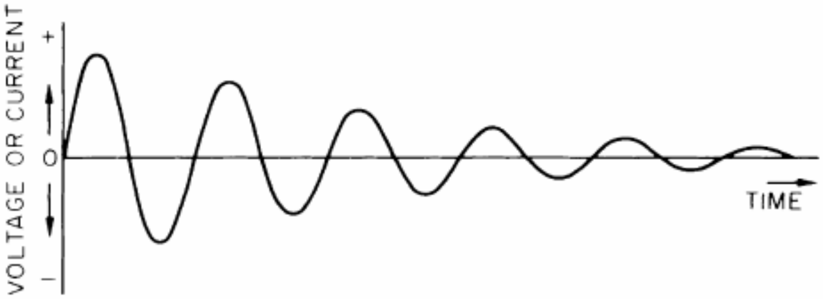


Fig. 14.3 The first few cycles of a train of damped oscillations

In Sec.8.4 the symbol Q was given to the ratio of reactance (of either kind) to resistance in a tuned circuit. So a highly damped circuit is a low-Q one. If the Q is as low as ½ the circuit is said to be critically damped; below that figure it is non-oscillatory and the voltage and current do not reverse after the initial kick due to an impulse but tend towards the shape shown in Figs. 3.6 and 4.10. A high-Q circuit continues oscillating for many cycles. Oscillation is created in any tuned circuit by any sudden electrical impulse, and the circuit is said to be shock-excited. The same kind of thing happens to a car when it encounters a hole or bump in the road, unless sufficient mechanical resistance is imparted to the springs in the form of shock absorbers (or dampers).

14.5 Maintaining Oscillation

What we want is a method of keeping an oscillator going by sufficient to cancel losses, oscillation is maintained *and the question of which electrode is common to input and output does not arise.* If we choose to regard c as the common point (as in Fig. 14.4 for example) then the input ca and the output cb need to be opposite in polarity, as they in fact are with common source (or emitter or cathode). But anyone is entitled to say that a is the common point and that this is a common-gate circuit, in which the amplified output ab has the same sign as the input (ac). A third party could equally maintain that this was a common-drain circuit, in which the output voltage bc is always not only of the same sign as the input (ba) but slightly smaller, and oscillation is equally well maintained on this basis, for different points of view of observers make no difference to the circuit!

14.6 Practical Oscillator Circuits

Basically, then, an oscillator consists of an oscillatory circuit (L and C in Fig. 14.6) and a maintaining amplifier (the rest of Fig. 14.6). Owing largely to the necessity for feeding the amplifier with d.c., practical oscillator circuits usually look very different from Fig. 14.6, and it is often quite difficult to see that in principle they are the same.

A resistive potential divider is rarely used, for it adds to the losses. And it is quite unnecessary; the oscillatory circuit can be made to act as its own potential divider. One method is to 'tap' the coil, as in Fig. 14.7. (To show no bias in favour of any particular amplifying device, here a bipolar transistor is used.) This is the basic form of what is called the Hartley circuit. Alternatively we can divide the oscillatory voltage in the C side. This could be done by inserting a third plate between the two, but usually it is more convenient to use two separate capacitors as in Fig. 14.8, of such capacitances that the two in series are equal to *C*. This is the Colpitts oscillator circuit.

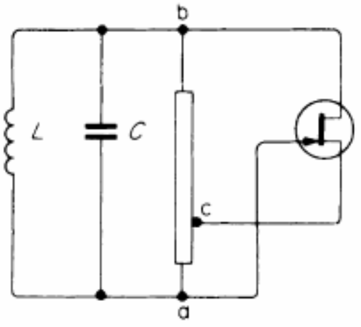


Fig. 14.6 A basic oscillator, with feeds omitted for clarity. Note that the question of which electrode of the amplifying device is the common one does not arise

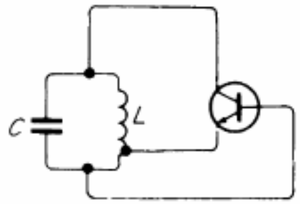


Fig. 14.7 This variety of Fig. 14.6 is the basis of the important Hartley oscillator circuit

connections to the LC circuit in order to obtain oscillation compared with the three (at least) required by the amplifier maintaining circuits discussed previously.

A disadvantage of circuit *a* is that both ends of the LC circuit are at r.f. potential with respect to earth so that it is impossible to earth the moving vanes of the tuning capacitor C, often a practical requirement. This difficulty can be overcome, as shown at *b*, by including in the emitter circuit an inductor with a high reactance at the oscillation frequency. This permits the emitter to be at r.f. potential and the base (connected to the moving vanes of C) to be earthed. The capacitance C_C is included to isolate L (one end of which is now earthed) from the positive potential of the collector. Because the lower end of L is at earth potential it is possible to introduce the positive supply voltage for the collector at this point and to eliminate the parallel-feed components R_C and C_C. The circuit diagram now appears as at *c*, and without showing the two fundamental inter-electrode capacitances this is very difficult to recognise as an oscillator circuit, let alone a Colpitts oscillator.

14.7 Resistance-Capacitance Oscillators

We have approached the subject of oscillators by considering an *LC* circuit with its natural frequency of oscillation, and then connecting an amplifier to feed it with regular periodical doses of energy to prevent those natural oscillations from dying out owing to circuit resistance or loading. We can alternatively regard the *LC* circuit simply as a positive-feedback path from output to input of the amplifier, so designed that only a.c. of one particular frequency is able to reach the input at the strength and phase relationship capable of maintaining itself via the amplifier. A circuit that selects or rejects particular frequencies or bands of frequencies is known as a *filter*. Filters are not necessarily oscillatory circuits, or even combinations of *L* and *C*. Fig. 6.10 shows how, when a voltage such as *ab* is applied to a capacitance and resistance in series, the voltages across C and R separately lead or lag *ab* by a phase angle determined by *C* and *R* and the frequency of the voltage. Now, as Fig. 14.5 showed, the input to a common-source (or cathode, or emitter) amplifier required to maintain oscillation must be inverted with respect to the output, and at least $1/A$ times the output in strength, A being the gain of the amplifier. Now for a sine wave inversion gives the same waveform as 180° phase shift. So a network which gives 180° phase shift at the required frequency of oscillation will also be satisfactory. Study of the situation illustrated by Fig. 6.10 shows that by the time the phase difference is 90° the 'output' voltage is nil. But by using three stages of this elementary form of filter, as in Fig. 14.14, a total phase difference of 180° can be achieved with an overall loss which can be matched by the gain of a transistor or valve. The output of this filter is connected to the input of the amplifier, whose output is connected to the input of the filter, as in Fig. 14.15.

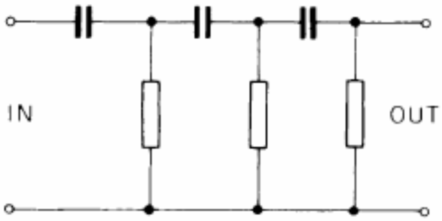
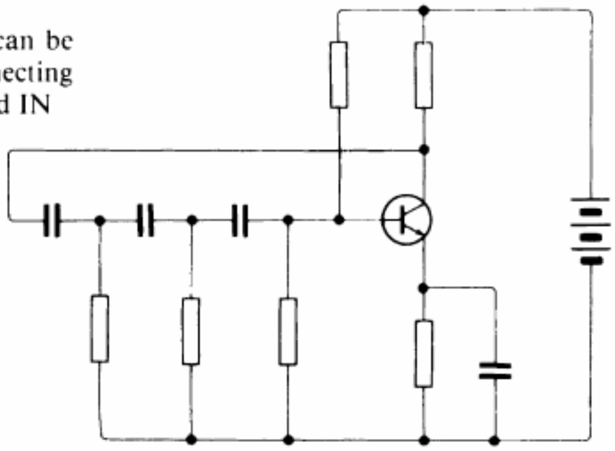


Fig. 14.14 Three-stage CR filter circuit for obtaining the 180° phase shift required by a self-oscillator

Fig. 14.15 The filter in Fig. 14.14 can be developed into an oscillator by connecting a simple amplifier between OUT and IN



This type of oscillator has a rather limited use, because the component values and the active device have to be well chosen to ensure oscillation; varying the frequency of oscillation is not very convenient because several component values have to be varied in order to do so; because of stray capacitances it is not really suitable for high frequencies; and the waveform is prone to depart from the sine shape. But it does illustrate the possibility of a sinusoidal oscillator without the inductance that is essential in a natural oscillatory circuit, and *RC* oscillators in more elaborate forms are very useful as audio-frequency signal generators.

14.8 Negative Resistance

Going back to the *LC* oscillatory circuit, we recall that the reason for the dying out of its oscillation was resistance. In Chapter 8 we studied the total effective resistance of such circuits and came to the conclusion (Sec. 8.16) that it could be represented as a (normally low) resistance in series with L and C, or as a (normally high) resistance in parallel with them. Fig. 14.16 may bring it to mind. Since then we have come across duality (Sec. 12.2) and can now see *a* and *b* as duals of one another and be inclined to refer to the high resistance as a low conductance. Now if we had such a thing as negative resistance we could insert $-r$ ohms in series in Fig. 14.16*a* could alternatively be simulated by the equivalent current generator in parallel with R (Sec. 12.1).

The tunnel diode oscillator is another example of an oscillator circuit which requires only two connections to the resonant circuit.

14.9 Amplitude of Oscillation

So far it has been asumed that the amplifier or other negative-resistance device used for maintaining oscillation was applied only just sufficiently to make good the losses. This is quite an unpractical condition, because the adjustment would be impossibly delicate in the first place, and even if it could be achieved it would be upset by the slightest change in temperature, component or device parameters, etc., so that oscillation would either start dying out or building up. In the latter event, what stops it growing so that it levels out? The answer is that the range of voltage or current over which the negative-resistance or amplifying characteristics exist is limited. This is very easily seen in Fig. 14.17 in which the extent of the negative-resistance region of the characteristic is limited to approximately 0.25 V. The peak-to-peak amplitude of the oscillation is limited to this value by the damping imposed on the tuned circuit by the positive-resistance regions below A and above B. Similarly we have already noted that the amplifying parameters we studied in Chapter 12 are valid only so far as the slopes of the characteristics at the working point continue linearly in both directions. This is why they were called small-signal parameters. Sooner or later those slopes come to an end, and with them the properties that made the amplitude of oscillation increase.

This can be illustrated by an example, Fig. 14.19. Although not identical with any of the previous oscillator circuits, all its features have already been explained. The I_C/V_C characteristics and load line are the same as in Fig. 13.2, and so is the working point P, but there is an interesting difference in the conditions represented thereby. In Fig. 13.2 the load was a real resistance, so of the supply voltage V_{CC}, which was 12V, half was lost in the resistance and only the remaining half reached the transistor as V_C at point P. In fig. 14.19 however the resistance of the coil L to d.c. is likely to be negligible, so only a 6 V battery is needed to maintain the working V_C at 6V.

The load line represents the dynamic resistance of the LC circuit, at the frequency of oscillation only (Sec. 8.10) and indicates by its slope that this is 6 kΩ. It is the parallel equivalent of the oscillation-frequency losses in L and C and also the resistance across the feedback coil multiplied by the square of the effective transformer step-up ratio from it to L (Sec. 7.11). Part of the resistance across the feedback coil is the input resistance of the transistor, which we know depends very much on the working point—see Fig. 10.12, for effect that is more easily controlled in this respect than the amplifying device itself.

Up to this point we have assumed that f.e.ts and valves are biased to a point that allows a range of oscillation between cut-off at one end and the start of grid or gate current at the other. Within this range practically no input current flows and the oscillatory circuit is not loaded thereby. But beyond it the amplitude of oscillation tends to be limited by output current being cut off, or the *LC* circuit being loaded by input current, or both.

14.10 Automatic Biasing of Oscillators

In many oscillator circuits, however, input current is allowed to flow and is used to generate bias for the active device. Fig. 14.20 shows the circuit diagram of a Hartley oscillator using this technique. The f.e.t. has zero gate bias, R_g being returned to the source terminal, and the transistor (assumed a depletion type) amplifies, so initiating oscillation in LC. As the amplitude of oscillation builds up, each positive peak at the gate causes the transistor to take a burst of gate current. This has two effects. Firstly it applies damping to the tuned circuit. Secondly it charges the capacitor C_g and the polarity of the charge biases the gate negatively. After the pulse of gate current has ceased, C_g begins to discharge through R_g and the voltage across C_g falls. The extent of the fall depends on the time constant R_gC_g and if this is long the fall is negligible before the next pulse of gate current restores the charge on C_g. In this way an almost steady voltage is maintained across C_g, and this provides the transistor with negative gate bias. Its value may be such that the transistor is cut off during negative peaks of oscillation at the gate, and takes a burst of gate current and drain current on positive peaks. The pulses of drain current provide the energy necessary to maintain the LC circuit in oscillation and the bursts of gate current abstract energy from it. The amplitude of oscillation thus stablises at a value at which the energy received from the drain circuit just balances the losses due to gate current and the inherent resistance in the LC circuit.

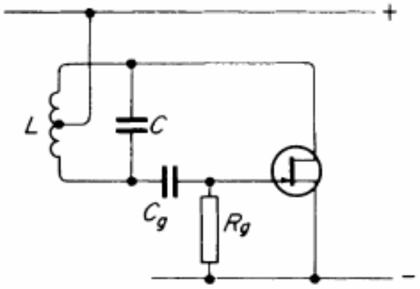


Fig. 14.20 A Hartley oscillator circuit using an f.e.t. and incorporating automatic gate bias

This biasing method can be used only with active devices such as valves, bipolar transistors and jugfets which take input current. It cannot be used with igfets which do not take input current.

14.11 Distortion of Oscillation

Section 14.9 can be summarised by saying that the amplitude of oscillation grows until it is limited by non-linearity of the characteristics of the maintaining device. As we shall be seeing more fully in Chapter 19, non-linearity causes distortion of the oscillation waveform, which in its natural state is sinusoidal. One cycle appears in full line in Fig. 14.21. Under the conditions shown in Fig. 14.19 both peaks of collector current would be cut off fairly sharply, as shown by the dashed line, representing severe distortion of the current waveform. But it would be wrong to suppose that this represents distortion of the oscillation waveform. To repeat: this is the waveform of the collector current. Its job is merely to make good the loss of energy in the oscillatory circuit through which it flows. Our study of this basic circuit in Chapter 8 showed that the oscillatory circulating current is Q times as much as needs to be supplied from outside. So no matter how badly this supplied current is distorted, it cannot affect the waveform of the oscillatory current much if Q is large. To go back to our clock analogy, the pendulum maintains its steady oscillation 'undistorted' although the periodical pushes administered by the escapement are anything but sinusoidal in 'waveform'. And it is quite usual for electronic oscillators to be kept going by square waves, pulses or what have you, as we have already seen in the last section.

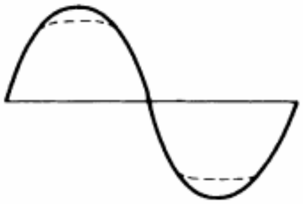


Fig. 14.21 If the amplitude of oscillation goes past the ends of the load line in Fig. 14.19 the peaks of collector current will be flattened as shown

But to obtain a very pure waveform—and this is often required of an oscillator—it is sensible to make sure that the oscillatory circuit itself has as high a Q as practicable, and that this is not spoilt by resistance transferred into it by connecting the maintainer. In other words, the maintainer should not be coupled more closely to the LC circuit than is necessary to ensure adequate oscillation in all conditions of use. The higher the Q the less maintenance is needed. The loose-coupling policy also tends to prevent even the small maintaining part of the current from being badly distorted.

For many purposes non-sinusoidal waveforms are required; methods of generating them will be deferred to Chapter 25.

14.12 Constancy of Frequency

The frequency of an LC oscillator, as we have seen, is adjusted by varying the inductance and/or capacitance forming the oscillatory circuit. For most purposes it is desirable that when the adjustment has been made the frequency shall remain perfectly constant. This means that the inductance and capacitance (and to a less extent the resistance) shall remain constant. For this purpose inductance includes not only that which is intentionally provided by the tuning coil, but also the inductance of its leads and the effects of any coils or pieces of metal (equivalent to single short-circuited turns) inductively coupled to it. Similarly capacitance includes that of the coil, the wiring and the maintaining device. Resistance depends on many things, such as loading and feed voltage variations. So some of the conditions for constant frequency are the same as those for undistorted waveform: high Q and minimum coupling with the maintaining device.

Inductance and capacitance depend mainly on dimensions, which expand with rising temperature. If the tuning components are shut up in a box along with the maintaining device, the temperature may rise considerably, and the frequency will drift for some time after switching on. Frequency stability therefore is aided by placing any components which develop heat well away from the tuning components; by arranging effective ventilation; and by designing tuning components so that expansion in one dimension offsets the effect of that in another. Fixed capacitors are obtainable employing a special ceramic dielectric material whose variations with temperature oppose those of other capacitors in the circuit; this reminds one of temperature compensation in watches and clocks.

An alternative method of achieving constant frequency is explained in the next section.

14.13 Crystal Oscillators

Oscillators for certain purposes require very good frequency stability. An example is the oscillator acting as the source of carrier frequency in a radio sender. The function of this, usually called the master oscillator, is to define the frequency of the station at least as accurately and constantly as the now very narrow international regulations require. Many such oscillators rely on an important aid to constant frequency which has not so far been mentioned. It exploits the remarkable properties of certain crystals, notably quartz. Such crystals are capable of vibrating mechanically up to high radio frequencies and in doing so they develop an alternating voltage between two opposite faces. Conversely, an a.v. applied between the faces causes vibration. This is known as the *piezo-electric* effect.

The subject is a large and specialised one, but for our purpose the crystal can be regarded as a tuned circuit of remarkably high inductance and Q. Unlike the coil-and-capacitor circuit, a crystal once it has been cut to the size and shape needed to vibrate at the required frequency has hardly anything about it that can vary. There is a small change in frequency with temperature, so where requirements are stringent the crystal is kept in a temperature-controlled box often called an oven.

The crystal is mounted between two metal plates, which are sometimes coated on to it, and for best results is enclosed in a vacuum. Fig. 14.22 shows an oscillator circuit in which the crystal itself is the only tuned circuit. The nature of this circuit is not clear unless the capacitances of the valve's anode and grid (shown dashed) are taken into account; it then turns out to be a Colpitts oscillator (Fig. 14.8). A disadvantage of this circuit is that most crystals are capable of oscillating at what are called spurious frequencies. To prevent this the anode coupling resistor is often replaced by an *LC* tuning circuit, which has insufficient impedance to sustain oscillations at any frequency other than the one to which it is tuned. Many other crystal-controlled oscillator circuits, both valve and transistor, have been devised.

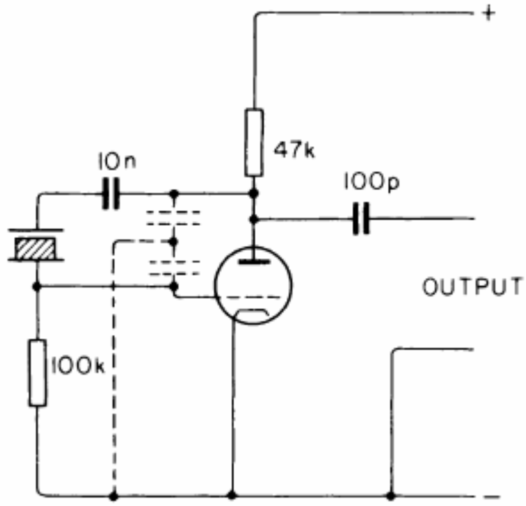


Fig. 14.22 One of the many types of crystal oscillator circuit

Although the frequency of a crystal-controlled oscillator can be adjusted within very narrow limits (a few parts in 10 000) by means of a variable capacitor in parallel, it is virtually a fixed-frequency oscillator. Crystals being relatively inexpensive, they can be used in a sender working on a number of spot frequencies. But if it is necessary to be able to vary the oscillator over a band of frequencies an *LC*-tuned variable frequency oscillator (v.f.o.) must be used, with appropriate precautions.

14.14 Microwave Oscillators

We saw in Sec. 8.17 that at microwave frequencies cavity resonators take the place of LC circuits, and we shall now see how such

CHAPTER 15

Modulation

15.1 The Need for Modulation

We have now reached the point—in theory at any rate—at which we can generate a continuous stream of waves of constant frequency. This, however, is completely useless for communication of messages, let alone speech, music or pictures. It merely establishes a link between sender and receiver, and to effect communication via the link some kind of organised change is needed. The technical word for it is *modulation.* The raw material that we have provided ready for modulation is called the *carrier wave*, although really it is a continuous succession of waves.

15.2 Amplitude Modulation

As a simple example of modulation suppose the 'programme' to be sent is a 1 kHz tuning note, which is heard as a high-pitched whistle. If it is to be a 'pure' sound its waveform must be sinusoidal, as in Fig. 15.1*a*. It can be generated by a suitable a.f. oscillator but, for the

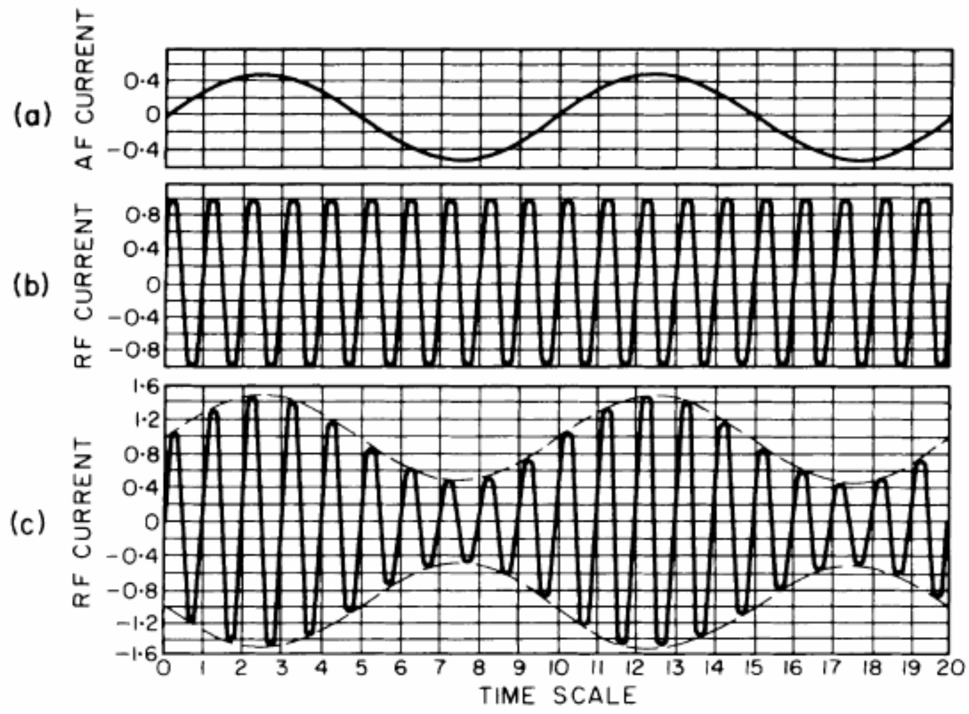


Fig. 15.1 Curve (*a*) represents two cycles of audio-frequency current, such as might be flowing in a microphone circuit; (*b*) represents the much larger number of radio-frequency cycles generated by the sender. When these have been amplitude-modulated by (*a*), the result is as shown at (*c*)

When the carrier has its maximum swing, from zero to double its mean value, it is said to be modulated to a depth of 100%. In general, half the variation in amplitude, maximum to minimum, expressed as a percentage of the mean amplitude, is taken as the measure of modulation depth. Thus, when a 100 V carrier wave is made to vary between 40 V and 160 V, the depth of modulation is 60%.

In transmitting a musical programme, variations in loudness of the received music are produced by variations in modulation depth, these producing corresponding changes in the amount of a.f. output from the receiver.

Similarly in transmitting a television programme the variations in light and shade occurring along each of the scanning lines are represented by the amplitude of the carrier wave, but the necessity for including synchronising signals and colour information complicates the modulation process as explained in Secs. 23.6 and 11.

15.3 Methods of Amplitude Modulation

In the early days, modulation was usually effected by varying the voltage of one of the electrodes of the oscillator valve by means of an amplified version of the sound, etc., to be communicated. This inevitably varied the frequency of oscillation to a small but no longer acceptable extent, so the practice now is to modulate the carrier wave in one of the r.f. amplifier stages. This is preferably not done in the first stage, in case it reacts on the oscillator, but the earlier in the chain it is done the less is the a.f. power needed for modulation. This advantage of what is called low-level modulation is offset by the need to avoid any distortion of the envelope from there on. That rules out Class C, for a start (Sec. 11.10). Class B r.f. amplification of a modulated carrier within tolerable limits of distortion is possible but difficult. So in spite of the very large amount of a.f. power needed, modulation is often applied at the final power amplifier stage.

Fig. 15.3 shows in outline the classical method of modulation, often called the Heising method, which varied the anode voltage. This common-cathode amplifier was fed with input signal from a tuned circuit L_1C_1, the output being developed in the tuned circuit L_0C_0. Both LC circuits were assumed tuned to the carrier frequency and, for sake of simplicity, various arrangements necessary in a practical version of this circuit have been omitted.

C had sufficient capacitance to provide a direct path for r.f. current to cathode but not enough to shunt the a.f. transformer T appreciably. This transformer coupled the output of an a.f. amplifier to the r.f. power amplifier. The amount of a.f. power needed was anything up to half as much as that supplied by the d.c. anode feed source, and of the same order as the r.f. output, so for a high-power sender the transformer had to be a very substantial one, and so had the a.f. amplifier feeding it. At 100% modulation its peak output voltage was almost equal to the feed d.v., and it therefore swung the total anode supply voltage from nearly zero to almost double the normal, thereby making the tuned circuit current vary as shown in Fig. 15.2.

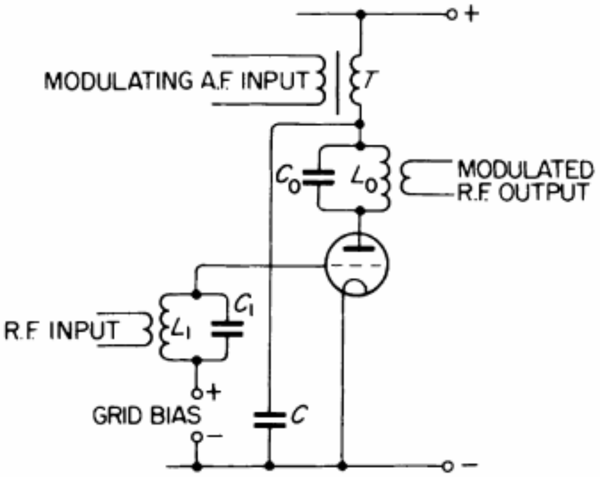


Fig. 15.3 Outline of Heising method of amplitude modulation

It used to be common practice to use a single inductor (a 'choke coil') in place of the transformer, and to connect the final a.f. amplifying valve across C, so that it was fed from the same source as the r.f. valve. This form of anode modulation was known as choke control.

Because of the large amount of power and heavy equipment needed for anode modulation, the a.f. voltage was sometimes applied to the grid. This was not so simple or satisfactory as might have been expected; it tended to increase distortion and certainly reduced the r.f. efficiency, so was not much favoured. Other methods of modulation have been devised, some of them very complicated, in attempts to reduce power consumption without increasing distortion.

15.4 Frequency Modulation

Instead of keeping the frequency constant and varying the amplitude, one can do vice versa. Fig. 15.4 shows the difference; *a* represents a small sample—two cycles—of the a.f. programme, as in Fig. 15.1*a*, and *b* is the corresponding sample of r.f. carrier wave after it has been amplitude-modulated by *a*. If, instead, it were frequency-modulated it would be as at *c*, in which the amplitude remains constant but the frequency of the waves increases and decreases as the a.f. voltage rises and falls. In fact, with a suitable vertical scale of hertz, *a* would represent the frequency of the carrier wave.

15.5 Telegraphy and Keying

To produce telegraph signals such as morse, a simple but extreme form of modulation is used: its depth is always either 0 or 100%. It used to be done by a morse key—a form of switch that makes a contact when pressed and breaks it when let go—in such a way that it turned the output of the sender on and off. Nowadays most of the remaining radiotelegraph communication is done at high speed by punched tape controlling contactors, which are electrically operated switches. There are a great many alternative positions in the sender circuits where a key or contactor would start and stop the output, but in practice quite a number of problems arise. So a.m. has been largely displaced by f.m., or frequency-shift keying, in which the radiation is going all the time but alternates between two closely spaced frequencies.

15.6 Methods of Frequency Modulation

The frequency of oscillation of an LC circuit can be altered by changing the value of the inductance or of the capacitance, and frequency modulation could therefore be achieved by using for the latter a component of which the capacitance can be controlled by the modulating signal. Such a component is a capacitor microphone because its capacitance varies in accordance with the sound waves striking it. Another possibility is the varactor diode, a current-controlled capacitor (Sec. 9.11). However these devices are unlikely to be used in practice because of the difficulty of achieving the required degree of linearity between frequency deviation and modulating signal, of controlling the extent of the deviation and of ensuring constancy in the unmodulated frequency (known as the centre frequency). These factors are important in f.m. senders and to achieve them circuits known as reactance modulators (Sec. 19.21) are used.

A simple method of frequency modulation is possible using a reflex klystron as the oscillator (Sec. 14.14) by applying the modulating signal to the repeller electrode. Microwave links sometimes use such a modulator.

15.7 Sources of Modulating Signal

So far we have learnt something about two of the boxes at the sending end of Fig. 1.11: the r.f. generator and the modulator.

A device by which sound, consisting of air waves, is made to set up electric currents or voltages of identical waveform, which in turn are used to control the modulator, is one that most people handle more or less frequently, in telephones. It is the *microphone.* There are a number of ways in which the variations in air pressure constituting a signals in the correct proportions but it is now common to incorporate a fourth tube in a colour camera specifically to generate the plain television signal; its light input is, of course, derived from the scene image before this enters the colour-splitting filter.

The television camera is not the only source of video signal in a television broadcasting system. Video tape reproducers are also used and video signals can also be obtained from cinema film using an equipment known as telecine. Television games also generate video signals.

So far we have confined our discussion to transmission of sound and vision signals as in broadcasting, but much of the information applies to the equipment used in radio telephones and by amateur broadcasters (radio hams). Senders have a host of other applications—military, in radio astronomy, in space travel, in Citizen's Band radio, by police, hospital services, fire brigades to name but a few. They are also used for remote metering purposes; e.g., to give information about the depth of water in a reservoir. This is an example of telemetry, and the modulating signal is likely to be in pulse form.

15.8 Theory of Sidebands

Let us now return to the waveform of an amplitude-modulated signal (Fig. 15.1*c*) and examine it in detail; in particular, to consider its frequency. Strictly speaking, it is only an exactly recurring signal that can be said to have one definite frequency. The continuous change in amplitude of a carrier wave during amplitude modulation makes the r.f. cycle of the modulated wave non-recurring, so that in acquiring its amplitude variations it has lost its constancy of frequency.

A mathematical analysis shows that if a carrier of f_1 Hz is modulated at a frequency f_2 Hz the modulated wave is exactly the same as the result of adding together three separate waves of frequencies f_1, (f_1+f_2), and (f_1-f_2). It is not easy to perform the analysis of the modulated wave into its three components by a graphical process, but the corresponding synthesis, adding together three separate waves, demands little more than patience.

Fig. 15.5 shows at *a*, *b*, and *c* three separate sine-wave trains, there being 35,30, and 25 complete cycles, respectively, within the width of the diagram. By adding the heights of these curves above and below their own zero lines point by point, the composite waveform at *d* is obtained. It has 30 peaks of varying amplitude, and the amplitude rises and falls five times in the period of time represented. If this is a thousandth part of a second, curve *d* represents what we have come to know as a 30 kHz carrier amplitude-modulated at 5 kHz. Fig.

A most important point to remember is that with either system of modulation the frequency band occupied depends entirely on the modulation and not at all on the frequency of the carrier wave.

15.9 Channel Separation

It is now clear that in order to receive a modulated signal completely the receiver must accept not one frequency but a whole band, preferably with equal effectiveness throughout. Outside this frequency band the receiver should ideally have no response, so that the sidebands of transmissions on nearby carriers are rejected. This is illustrated for a.m. in Fig. 15.8 in which the dashed line shows an ideal overall receiver response curve embracing the whole of transmission A and completely rejecting transmission B. This curve illustrates the selectivity of the receiver; i.e., its ability to accept a wanted transmission and to reject those on neighbouring frequencies.

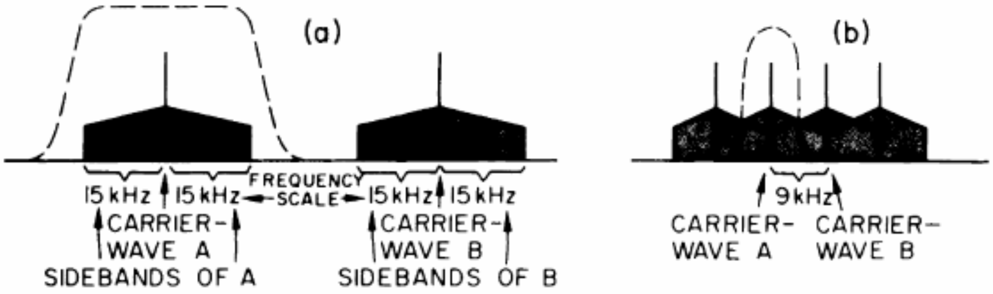


Fig. 15.8 At (*a*) is shown how medium-wave broadcasting stations' carrier waves ought to be spaced on the frequency scale in order to permit full reception of each without interference from others. How they actually are spaced (*b*), shows that interference is almost inevitable if full reception of sidebands is attempted

For sound, the modulation frequencies are audio frequencies, which go up to at least 15 kHz. The corresponding sidebands would occupy at least 30 kHz. Ideally, then, the carrier waves of a.m. sound broadcasting stations should be spaced at intervals of not less than 30 kHz, plus a margin for selectivity, as suggested in Fig. 15.8*a*.

Unfortunately the pressure of national demands for broadcasting 'channels' has squeezed this ideal severely; the standard interval in Europe on the low and medium frequencies is only 9 kHz. Obviously this faces receivers with an impossible task, for it brings 9 kHz modulation of one station on to the same frequency as the carrier wave of the next and the 9 kHz modulation of the next but one. Therefore few, if any, broadcasting stations on these frequencies actually radiate sidebands up to 15 kHz; 8 kHz is a more realistic maximum. There is therefore some loss at the start. And even what is transmitted cannot be fully reproduced without interference, unless the wanted transmission is so much stronger than those on adjacent frequencies as to drown them. So receivers for other than local-

15.10 Multiplex

We have so far discussed methods of modulating a carrier by a single modulating signal. But in stereo radio and colour television broadcasting, a second modulating signal is needed (as explained in later chapters) and it would be very convenient if the same carrier could be used to carry both, provided that the signals could be recovered without mutual interference at the receiver.

This can be achieved, and the same method is used in both radio and television. The second signal is used to amplitude-modulate a second carrier, known as a subcarrier, which is added to the first signal, the frequency of the subcarrier being so chosen that its sidebands do not interfere with those of the first signal. In the Zenith-GE stereo system used in the UK, the subcarrier is at 38 kHz and its sidebands occupy the range 23-53 kHz. The sidebands of the first signal (known as a baseband signal because it begins near z.f.) extend to 15 kHz so that there is a guard band of 8 kHz between the two sets of sidebands. The composite signal is used to modulate the main carrier. Receivers intended for monophonic sound reproduction or plain television ignore the subcarrier and its modulation.

This technique can be extended to embrace any number of signals. By choosing a different carrier frequency for each signal, the sidebands can be made to occupy a band without interference in the same way as radio channels are arranged within a waveband. This is known as frequency-division multiplex (f.d.m.) and is exploited in telephone systems to enable up to 900 telephone calls to be transmitted simultaneously over a single link.

15.11 Pulse Modulation Systems

Extensive use is made in modern electronics of pulses, which are waves of approximately rectangular shape. A train of similar pulses can be used to modulate a sinusoidal carrier, and can themselves be modulated to transmit information. For example the amplitude or the frequency of the pulses can be modulated as in the a.m. and f.m. systems described above, but a pulse signal permits other modulation methods not possible with a sinusoidal carrier such as pulse duration and pulse position modulation. These systems, apart from pulse amplitude modulation, employ trains of constant-amplitude pulses and these have a great advantage over sinusoidal systems. If a received pulse signal is weak it is likely to be almost obscured by the noise inevitably generated by the active and passive components in any transmission system. But, provided their amplitude exceeds that of the noise, the pulses can be used to trigger a pulse generator (such as a multivibrator, Sec. 25.5) which will produce much stronger constant-amplitude noise-free pulses with

CHAPTER 16

Transmission Lines

16.1 Feeders

In dealing with circuit make-up we have for the most part assumed that resistance, inductance and capacitance are concentrated separately in particular components, such as resistors, inductors and capacitors. True, any actual inductor (for example) has some of all three of these features, but it can be imitated fairly accurately by an imaginary circuit consisting of separate *L, C* and *R* (Sec. 8.16). In the next chapter, however, we shall turn attention to antennas, in which *L, C* and *R* are all mixed up and distributed over a considerable length. Antenna calculations are therefore generally much more difficult than anything we have attempted, especially as antennas can take so many different forms.

But one example of distributed-circuit impedance does lend itself to rather simpler treatment, and makes a good introduction to the subject, as well as being quite important from a practical point of view. It is the *transmission line.* The most familiar example is the cable connecting the television antenna on the roof to the set indoors. For reasons just explained (Sec. 15.9) television uses signals of very high frequency, which are best received on antennas of a definite length (Sec. 17.7). It is desirable that the wires connecting antenna to receiver should not themselves radiate or respond to radiation, nor should they weaken the v.h.f. or u.h.f. signals. The corresponding problem at the sending end is even more important, because a large amount of power has to be conveyed from the sender on the ground to an antenna hundreds of feet up in the air. A transmission line for such purposes is usually called a *feeder.*

To reduce radiation and pick-up to a minimum, its two conductors must be run very close together, so that effects on or from one cancel out those of the other. One form is the parallel-wire feeder (Fig. 16.1(*a*)); still better is the coaxial feeder (*b*), in which one of the conductors totally encloses the other.

16.2 Circuit Equivalent of a Line

With either type, and especially the coaxial, the closeness of the two leads causes a large capacitance between them; and at very high frequencies that means a low impedance, which, it might be supposed, would more or less short-circuit a long feeder, so that very distributed, every 10 metres of line will absorb a certain percentage of the power put into it. Obviously, a line made of a very thin wire, of comparatively high resistance, will cause greater loss than one made of low-resistance wire. So, too, will a pair separated by poor insulating material (Sec. 8.15). Although the loss due to resistance is an important property of a line, it is generally allowable to neglect the resistance itself in calculations such as we have just been making.

16.3 Characteristic Resistance

The next thing to consider is how to make the feeder fit the load, so as to fulfil the necessary condition, $\sqrt{(L/C)} = R$. $\sqrt{(L/C)}$ is obviously a characteristic of the line itself, and to instruments connected to one end the line appears to be a resistance; so for any particular line it is called its *characteristic resistance,* denoted by R_0. (If loss resistance is taken into account the expression is slightly more complicated than $\sqrt{(L/C)}$, and includes reactance; so the more comprehensive and strictly accurate term is characteristic (or surge) impedance, Z_0. With reasonably low-loss lines there is not much difference.)

If you feel that the resistance in question really belongs to the terminating load and not the line, then you should remember that the load resistance can always be replaced by another length of line, so that ultimately R_0 can be defined as the input resistance of an infinitely long line. Alternatively (and more practically) it is a resistance equal to whatever load resistance can be fed through any length of the line without making any difference to the generator.

In any case, L and C depend entirely on the spacing and diameters of the wires or tubes of which the line consists, and on the permittivity and permeability of the spacing materials. So far as possible, to avoid losses, feeders are air-spaced. The closer the spacing, the higher is C (as one would expect) and the smaller is L (because the magnetic field set up by the current in one wire is nearly cancelled by the field of the returning current in the other). So a closely-spaced line has a small R_0, suitable for feeding low-resistance loads. Other things being equal, one would expect a coaxial line to have a greater C, and therefore lower R_0, than a parallel-wire line. Formulae have been worked out for calculating R_0; for a parallel-wire line (Fig. 16.4*a*) it is practically $276 \log_{10}(2D/d)$ so long as D is at least 4 or 5 times d, and for a coaxial line (Fig. 16.4*b*) $138 \log_{10}(D/d)$. Graphs such as these can be used to find the correct dimensions for a feeder to fit a load of any resistance, within the limits of practical construction. It has been found that the feeder dimensions causing least loss make R_0 equal to about 600 Ω for parallel-wire and 80 Ω for coaxial types; but a reasonably efficient feeder of practical dimensions can be made to have an R_0 of anything from, say 200 to 650 and 40 to 140 Ω respectively. We shall see later what to do if the load resistance cannot be matched by any available feeder.

In the meantime, what if the load is not a pure resistance? Whatever it is, it can be reckoned as a resistance in parallel with a reactance. And this reactance can always be neutralized or tuned out by connecting in parallel with it an equal reactance of the opposite kind (Sec. 8.8). So that little problem is soon solved for any desired frequency.

16.4 Waves along a Line

Although interposing a loss-free line (having the correct R_0) between a generator and its load does not affect the voltage and current at the load terminals—they are still the same as at the generator—it does affect their phase. The mere fact that the voltage across R in Fig. 16.3*b* is equal to the voltage across the terminals makes that inevitable, as one can soon see by drawing phasor diagrams. Fig. 16.5*a* shows a few sections as in Fig. 16.3, and the phasor diagrams (*b*) are drawn by beginning at the load R and working backwards. They take the form of continuously opening fans. We can easily judge from this that for an actual line, in which the sections are infinitesimal, the diagrams would be circles. If we consider what happens in a line when the generator starts generating (Fig. 16.6) we can see that what the gradual phase lag along it really means is that the current takes time to travel from generator to load.

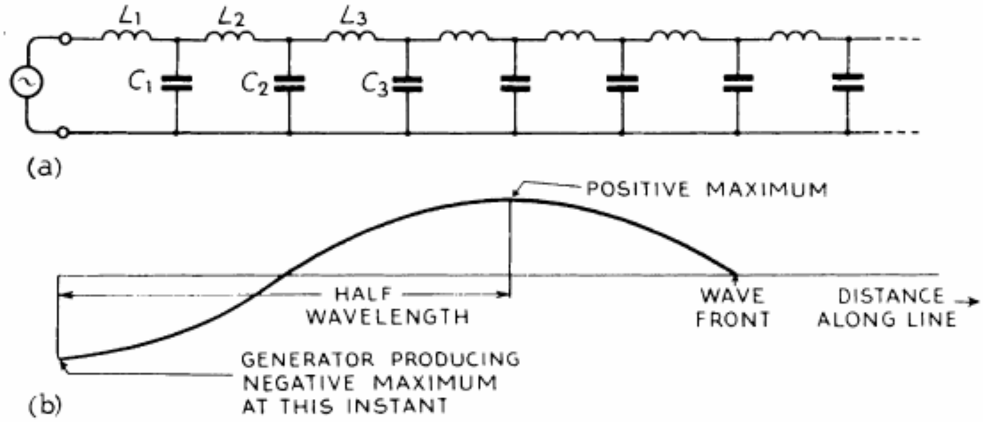


Fig. 16.6 (*a*) Electrical representation of the generator end of a transmission line, and (*b*) instantaneous voltage (and current) diagram $\frac{3}{4}$ of a cycle from the start of generating a sine wave

During the first half-cycle (positive, say) current starts flowing into C_1, charging it up. The inductance L_1 prevents the voltage across C_1 rising to its maximum until a little later than the generator voltage maximum. The inertia of L_2 to the growth of current through it allows the charge to build up in C_1 a little before C_2, and so on. Gradually the charge builds up on each bit of the line in turn. Meanwhile, the generator has gone on to the negative half-cycle, and this follows its predecessor down the line. At its first negative

Neglecting line loss, the voltage and current reaching the load end will be the same as at the generator. So if the load turns out to be a resistance equal to R_0, it will satisfy Ohm's law, and the whole of the power will be absorbed just as fast as it arrives. For example, if a piece of line having an R_0 of 500 Ω is connected to a 1000 V generator having an r of 500 Ω, the terminal voltage is bound to be 500 V and the current 1 A until the wave reaches the far end, no matter what may be there. If there is a load resistance of 500 Ω, the current will, of course, continue at 1 A everywhere.

16.5 Wave Reflection

But suppose that the load resistance is, say, 2000 Ω. According to Ohm's law it is impossible for 500 V applied across 2000 Ω to cause a current of 1 A to flow. Yet 1 A is arriving. What does it do? Part of it, having nowhere to go, starts back for home. More scientifically, it is *reflected* by the mismatch. The reflected current, travelling in the opposite direction, can be regarded as opposite in phase to that arriving, giving a total which is less than 1 A. The comparatively high resistance causes the voltage across it to build up above 500; this increase can be regarded as an in-phase reflected voltage driving the reflected current. If half the current were reflected, leaving 0.5 A to

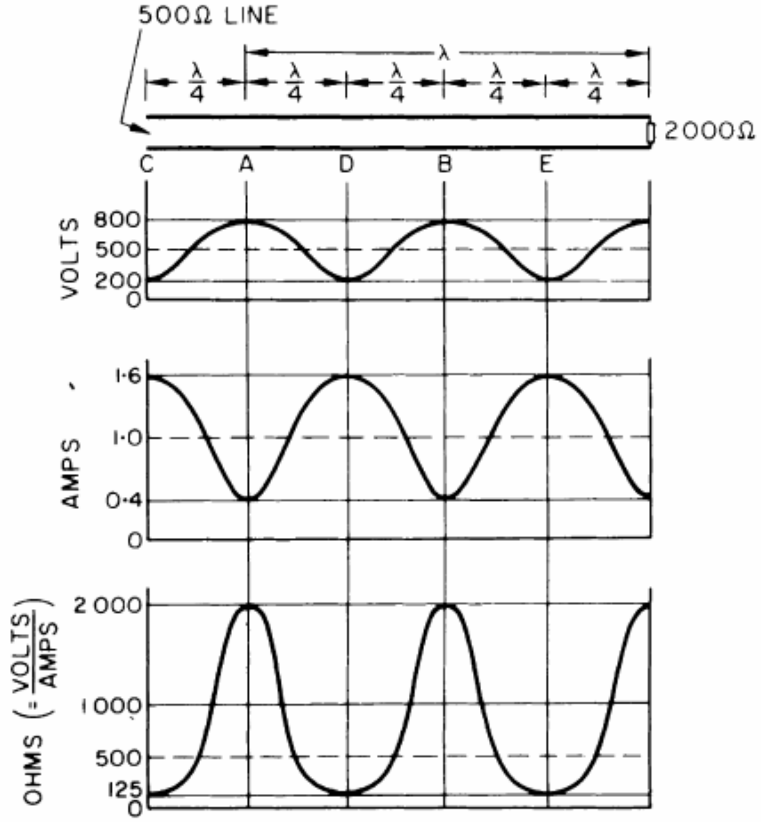


Fig. 16.7 Diagrams of voltage, current, and (as a result) impedance at the load end of an unmatched transmission line, showing standing waves

go into the load resistance, the voltage would be increased by a half, making it 750. A voltage of 750 and current 0.5 A would fit a 1500 Ω load, but not 2000 Ω, so the reflected proportion has to be greater—actually 60%, giving 800 V and 0.4 A at the load. In general terms, if the load resistance is R the fraction of current and voltage reflected, called the *reflection coefficient* or *return loss* and denoted by ρ, is $(R-R_0)/(R+R_0)$.

We now have 1 A, driven by 500 V, travelling from generator to load, and 0.6 A, driven by 300 V, returning to the generator. (The ratio of the reflected voltage to the reflected current must, of course, equal R_0.) The combination of these two at the terminals of the load gives, as we have seen, $1-0.6=0.4$ A, at $500+300=800$ V. But at other points on the line we have to take account of the phase lag. At a distance from the load end equal to quarter of a wavelength ($\lambda/4$) the arriving and returning waves differ in phase by half a wavelength ($\lambda/2$) or 180° as compared with their relative phases at the load (because a return jouney has to be made over the $\lambda/4$ distance). So at this point the current is $1+0.6=1.6$ A at $500-300=200$ V. At a point $\lambda/8$ from the load end the phase separation is $\lambda/4$ or 90°, giving 1.16 A at 582 V calculated as explained in Sec. 5.8. At intervals of $\lambda/2$, the two waves come into step again.

16.6 Standing Waves

Calculating the current and voltage point by point in this way and plotting them, we get the curves in Fig. 16.7. It is important to realize that these are not, as it were, flashlight photographs of the waves travelling along the line; these are r.m.s. values set up continuously at the points shown, and would be indicated by meters connected in or across the lines at those points (assuming the meters did not appreciably affect the R_0 of the line where connected). Because this wavelike distribution of current and voltage, resulting from the addition of the arriving and reflected waves travelling in opposite directions, is stationary, it is called a *standing wave.* For comparison, the uniform distribution of current and voltage when the load resistance is equal to R_0 is shown dashed.

The ratio of maximum to minimum current or voltage is called the *standing wave ratio* (s.w.r.); in our example it is $800/200=4$. (Because the ratio is the same for voltage and current the common practice of referring to voltage standing wave ratio (v.s.w.r.) is unnecessary and liable to mislead one into thinking that the ratio must be different for current.) Applying simple algebra to the way we have found the s.w.r. in this example we can easily find that in general it is equal to R/R_0. Its usual symbol is S.

In due course the reflected wave reaches the generator. It is in this indirect way that the load makes itself felt by the generator. If the

2000 Ω load had been directly connected to the generator terminals, the current would have been $1000/(500+2000)=0.4$ A, and the terminal voltage $0.4\times 2000=800$ V, and the power $800\times 0.4=320$ W. This, as we have seen, is exactly what the load at the end of the line is actually getting. But the power that originally went out from the generator, being determined by R_0, was $500\times 1=500$ W. The reflected power is $300\times 0.6=180$ W, so the net outgoing power is $500-180=320$ W, just as it would have been with the load directly connected.

16.7 Line Impedance Variations

So the power adjustment is (in this case) quite simple. But the current and voltage situation at the generator is complicated by the time lag, and is not necessarily the same as at the load. Unless the length of the line is an exact multiple of $\lambda/2$, the phase relationships are different. Suppose, for example, it is an odd multiple of $\lambda/4$; say, $5\lambda/4$, as in Fig. 16.7. Then the current and voltage at the generator end will be 1.6 A and 200 V respectively. That makes 320 W all right—but compare the power loss in the generator. 1.6 A flowing through $r=500\,\Omega$ is $1.6^2\times 500=1280$ W, whereas at an 800 V point the current would be only 0.4 A and the loss in the generator only $0.4^2\times 500=80$ W! So when there are standing waves, the exact length of the line (in wavelengths) is obviously very important. If there are no standing waves, it does not matter if the line is a little longer or shorter or the wavelength is altered slightly; and that is one very good reason for matching the load to the line. The usual way of making sure that the load is right is by running a voltmeter or other indicator along the line and seeing that the reading is the same everywhere, except perhaps for a slight gradual change due to line loss.

Connecting the generator to a point on the line where the voltage and current are 200 V and 1.6 A respectively is equivalent to connecting it to a load of $200/1.6=125\,\Omega$. The impedance of the line at all points can easily be derived from the voltage and current curves, as at the foot of Fig. 16.7. This curve indicates the impedance measured at any point when all the line to the left of that point is removed. The impedance depends, of course, not only on the distance along the line to the load but also on R_0 and the load impedance. From a consideration of the travelling waves it can be seen that at $\lambda/4$ intervals the phases of the currents and voltages are exactly the same as at the load, or exactly opposite; so that if the load is resistive the input resistance of the line is also resistive. At all other points the phases are such as to be equivalent to introducing reactance.

In our example we made the generator resistance, r, equal to R_0. If it were not, the situation would be more complicated still, because the reflected wave would not be completely absorbed by the generator, so part would be reflected back to the load, and so on, the reflected power being smaller on each successive journey, until finally becoming negligible. The standing waves would be the resultant of all these travelling waves. Even quite a long line settles down to a steady state in a fraction of a second, but it is a state that is generally undesirable because it means that much of the power is being dissipated in the line instead of being delivered to the load.

It should be noted that mismatching at the generator end does not affect the standing-wave ratio, but does affect the values of current and voltage attained.

Another result of making $r = R_0$ is that any reflection from the load or elsewhere inevitably impairs the generator-to-line matching. That is so even if the generator is connected to a point where the impedance curve coincides with the 500-Ω line in Fig. 16.7, because then the impedance is reactive. But if r were, say, 2000 Ω, so that it would be mismatched to the 500-Ω line during the brief moment following the start, there would be a chance that when the reflected wave arrived it would actually improve the matching, even to making it perfect (e.g., if connected at A or B in Fig. 16.7); but on the other hand it might make it still worse (if connected at C, D, or E).

It should be remembered that 'perfect' matching is that which enables maximum power to be transferred; but in practice there may be good reasons for deliberately mismatching at the generator. To take figures we have already had, it may be considered better to deliver 320 W with a loss of 80 W than the maximum (500 W) with a loss of 500 W.

16.8 The Quarter-Wave Transformer

An interesting result of the principles exemplified in Fig. 16.7 is that a generator of one impedance can be perfectly matched to a load of another by suitably choosing the points of connection. For instance, a generator with an internal resistance of 125 Ω would be matched to the 2000 Ω load if connected at C, D or E. The line then behaves as a 1:4 transformer. Points can be selected giving (in this case) any ratio between 1:4 and 4:1, but except at the lettered points there is reactance to be tuned out.

It is not necessary to use the whole of a long line as a matching transformer; in fact, owing to the standing waves by which it operates it is generally undesirable to do so. We can see from Fig. 16.7 that the maximum ratio of transformation, combined with non-reactive impedance at both ends, is given by a section of line only quarter of a wavelength long. We also see that the mismatch ratio to the line is the same at both ends; in this particular case it is

1:4 (125 Ω to 500 Ω at the generator end and 500 Ω to 2000 Ω at the load end), which, incidentally, is equal to the voltage ratio of the whole transformer. In general terms, if R_I denotes the input resistance of the line (with load connected), R_I is to R_0 as R_0 is to R; which means $R_I/R_0 = R_0/R$, or $R_0 = \surd(R_I R)$.

This formula enables us to find the characteristic resistance of the quarter-wave line needed to match two unequal impedances, R_I and R. Suppose we wished to connect a 70 Ω load to a 280 Ω parallel-wire feeder without reflection at the junction. They could be matched by interposing a section of line λ/4 in length, spaced to give a R_0 of $\surd(70 \times 280) = 140\,\Omega$. Parallel wires would have to be excessively close (Fig. 16.4*a*), but the problem could be solved by using a length of 70-Ω coaxial cable for each limb and joining the metal sheaths together as in Fig. 16.8, putting the 70 Ω impedances in series across the ends of the 280 Ω line.

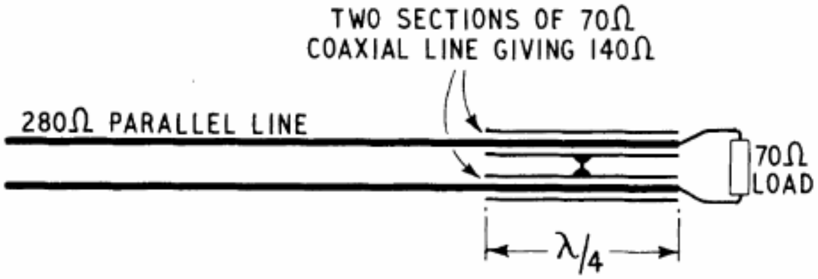


Fig. 16.8 Example of a quarter-wavelength of line being used as a matching transformer

16.9 Fully Resonant Lines

Going now to extremes of mismatch, we inquire what happens when the 'load' resistance is either infinite or zero; in other words, when the end of the line is open-circuited or short-circuited. Take the open circuit first. If this were done to our original 500 Ω example, the current at the end would obviously be nil, and the voltage would rise to 1000—double its amount across the matched load. This condition would be duplicated by the standing waves at points A and B in Fig. 16.7; while at C, D, and E the voltage would be nil and the current 2 A. The impedance curve would consequently fluctuate between zero and infinity.

With a short-circuited line, there could be no volts across the end, but the current would be 2 A; in fact, exactly as at E with the open line. A shorted line, then, is the same as an open line shifted quarter of a wavelength along. Reflection in both cases is complete, because there is no load resistance to absorb any of the power.

If the generator resistance is very large or very small, *nearly* all the reflected wave will itself be reflected back, and so on, so that if the line is of such a length that the voltage and current maximum points coincide with every reflection, the voltages and currents will build up to high values at those maximum points—dangerously high with a powerful sender. This reminds us of the behaviour of a high-Q resonant circuit (Sec. 8.3). The maximum current or voltage points are called *antinodes,* and the points where there is *no* current or voltage are *nodes.*

When the length of a short-circuited or open-circuited line is a whole number of quarter-wavelengths, the input impedance is approximately zero or infinity. An *odd* number of quarter-wavelengths gives opposites at the ends—infinite resistance if the other end is shorted, and vice versa. An *even* number of quarter-wavelengths gives the same at each end.

In between, as there is now no load resistance, the impedance is a pure reactance. At each side of a node or antinode there are opposite reactances—inductive and capacitive. If it is a current node, the reactance at a short distance each side is very large; if a voltage node, very low. Fig. 16.9 shows how it varies. It is clear from this that a short length of line—less than quarter of a wavelength—can be used to provide any value of inductance or capacitance. For very short wavelengths, this form is generally more convenient than the usual inductor or capacitor, and under the name of a *stub* if often used for balancing out undesired reactance.

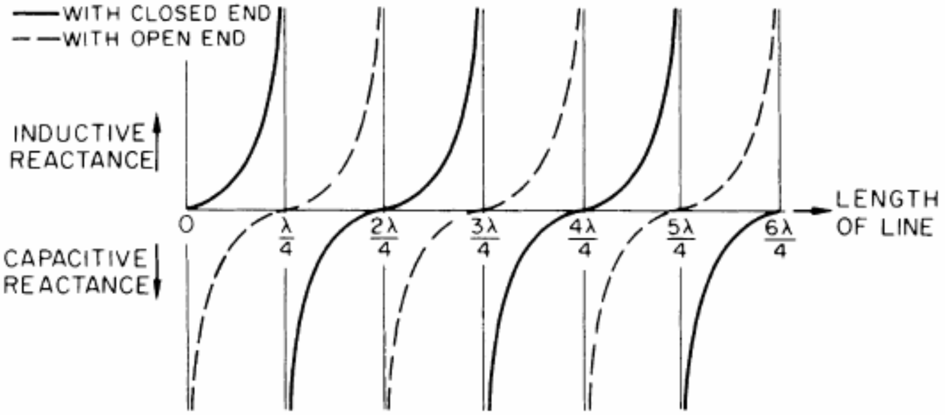


Fig. 16.9 Showing how the reactances of short-circuited and open-circuited lines vary with their length

16.10 Lines as Tuned Circuits

A quarter-wave line shorted at one end is, as we have just seen, a high—almost infinite—resistance. But only at the frequency that makes it quarter of a wavelength (or an odd multiple of $\lambda/4$). This resistance is closely analogous to the dynamic resistance of a parallel resonant circuit; and in fact a line can be used as a tuned circuit. At wavelengths less than about 1 metre it is generally more efficient and easily constructed than the conventional sort.

The lower the loss resistance of the wire the higher the dynamic resistance across the open ends; and to match lower impedances all that is necessary is to tap it down, just as if it were any other sort of tuned circuit.

If the line is made slightly shorter than $\lambda/4$ the open end behaves as an inductor and can be tuned by connecting a variable capacitor across it. Fig. 16.10*a* is an example of a conventional tuning circuit, such as might be used in a receiver, and *b* is the coaxial equivalent. A coaxial feeder is also used to connect the antenna, and being normally about 80 Ω it is tapped low down, near the earthed end. The impedance at the top end is normally thousands of ohms, and is too high for the input of a transistor so this too is tapped down.

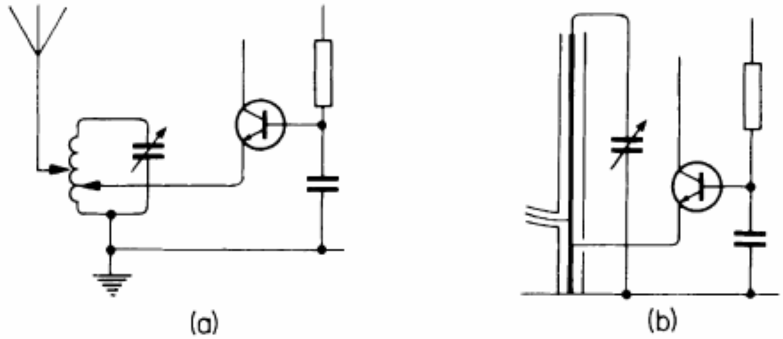


Fig. 16.10 (*a*) Conventional aerial input circuit using 'lumped' components, and (*b*) distributed equivalent using a quarter-wave coaxial line

This is in essence the type of circuit used in most u.h.f. tuners in television receivers, but the variable capacitor is usually replaced by a varactor diode. Sometimes it is more convenient to make the short-circuited line slightly less than $3\lambda/4$ long, or alternatively an open-circuited line of just under $\lambda/2$ may be used. Both give the required inductive impedance but the $\lambda/2$ line has the advantage that the centre conductor — being open-circuited — can be used to supply the tuning control current to the varactor diode. See Sec. 22.8.

16.11 Waveguides or Radio-wave Plumbing

Above approximately 200 MHz the loss by radiation from a twin-wire transmission line becomes excessive, but coaxial lines can be successfully used up to 3000 MHz. At still higher frequencies a different mode of propagation of radio waves is possible in a tube without an inner conductor, in which the wave progresses along the tube in a series of internal reflections at the tube walls. In this way a radio wave can be guided along a tube as water passes along a pipe. Such tubes, often of rectangular cross-section, are termed waveguides and are extensively used in radar. For successful transmission the wavelength must be less than twice the diameter of

CHAPTER 17

Radiation and Antennas

17.1 Bridging Space

In the first chapter of this book the processes of radio communication were traced very briefly from start to finish. Before going on to the receiver we should know something about how the space in between is bridged.

We have just seen how electrical power can be transmitted from one place to another at the not inconsiderable speed of nearly 300 000 km per second, in the form of waves along a line consisting of two closely spaced conductors. A southerner who was asked to explain how it was possible to signal by telegraph from New York to New Orleans replied by pointing out that when you tread on the tail of a dog the bark comes out at its head. A telegraph was like a very long dog, with the tail in New York and the head in New Orleans. 'But what about wireless?' he was asked. 'Same thing, 'cept the dawg am imagin'ry'. This reply, though ingenious, is not perhaps altogether satisfying. How is it possible to transmit the waves without any conductors to carry the currents to and fro?

The significant thing about the section on how waves are transmitted along a line (Sec. 16.4) is that conduction was hardly mentioned; the important properties of a line are its distributed capacitance and inductance. In other words, transmission of waves along a line is primarily a matter of oscillating electric and magnetic fields, located in the space between and around the two conductors. These conductors are not required to carry current from one end to the other, any more than waves on a pond carry the water from one side to the other; all that the electrons in the conductors do is oscillate a very short distance locally.

17.2 The Quarter-Wave Resonator Again

Near the end of the same chapter we found that a line only quarter of a wavelength long and open at the end, as in Fig. 17.1*a*, is equivalent from the generator's point of view to a series resonant circuit *b*; that is to say, the only impedance the generator faces is the incidental resistance of the circuit. We saw in Sec. 8.4 that if this resistance is small a very little generator voltage can cause a very large current to flow and build up large voltages across C and L. This remarkable behaviour of C and L is due to the electric field concentrated mainly between the plates of C and the magnetic field concentrated mainly so increasing L. The lines of electric field are on the average much longer, so C is decreased. Again, these two effects almost exactly cancel out, so the line is still nearly resonant to the same generator frequency. It can be brought exactly into tune by shortening it a very few per cent, which will be assumed to be done even though (as in Fig. 17.2) the total length is for simplicity marked $\lambda/2$. So the distribution of current and voltage along the conductors is much the same as before. Looking at Fig. 17.3 we would say that the current was greatest at the middle because it has to charge all the capacitance in parallel, whereas near the ends there is hardly any to charge.

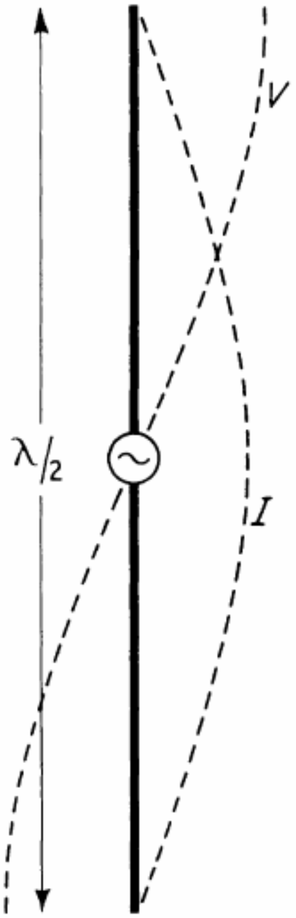


Fig. 17.2 The quarter-wave line of Fig. 17.1 opened fully out

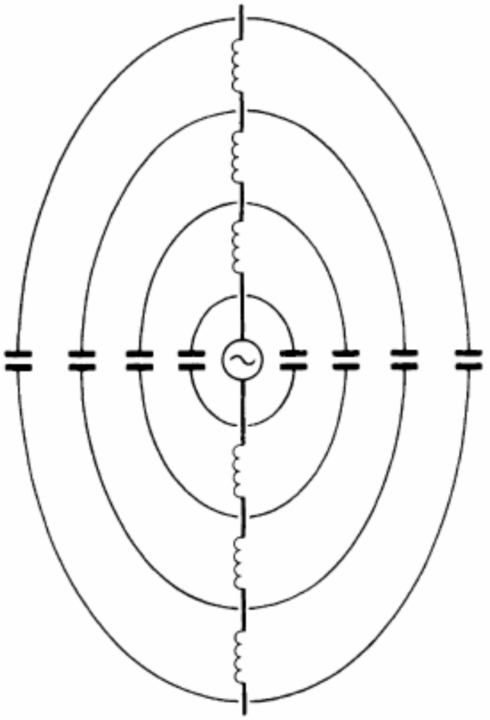


Fig. 17.3 Approximate representation of the distributed inductance and capacitance in the half-wavelength system of Fig. 17.2

Fig. 17.4 shows some typical electric and magnetic lines of force. At any one radius from the generator they would trace out a sort of terrestrial globe, on which the magnetic lines were the parallels of latitude and the electric lines the meridians of longitude. This is helpful as a picture, but we must remember once again that field lines are as imaginary as lines of latitude and longitude.

17.3 A Rope Trick

It is clear that quite an appreciable portion of each field is now far enough away to be subject to a phase delay in its reaction on the rod caused by the generator. When the current is upwards it charges the upper half positively and the lower half negatively, resulting in an electric field downwards between the two. Half a cycle later the pattern is reversed. The up-and-down displacement of the rope from its middle position represents this alternating electric field. The displacement travels outward from the hand as each bit of rope passes it on to the next. If the time of each cycle of hand movement is long compared with the time for the movement to pass along the rope, all the rope for at least quite a distance appears to move as a whole 'in phase'. Because the electric field has to spread out in all directions, and not just in one direction like the rope, its amplitude falls off much more rapidly with distance from the source than does that of the rope.

When the hand is moved so fast that the rope displacement has not gone far before the hand reverses, the displacements become progressively out of phase with one another, and this is why travelling waves are formed, as in Fig. 17.5*b*. The same applies to the alternations of electric field; they form a wavelike pattern which travels outwards, just like the electric field between the conductors of a long line in that direction, except for a falling off in amplitude because it is not going out only along one radius but all around the 'equator'.

17.4 Electromagnetic Waves

A good question at this stage would be: what keeps it going? The rope wave would not keep going unless the rope had not only displacement but *inertia.* The fact that we have already noted a resemblance between inertia and inductance (Sec. 4.9) may remind us that we seem to have forgotten about the magnetic field. But of course that too is spreading out, and in fact is essential to the wave motion. When we did the experiments with iron filings around a wire carrying current (Sec. 4.1) we said that the cause of the magnetic field was the electric current in the wire. But we know that an electric current consists of electric charges in motion, and charges are surrounded by an electric field. So the electric current, which generates the magnetic field, is really a moving electric field. We know, too, that a moving magnetic field generates an e.m.f. We tend to think, perhaps, of an e.m.f. only in a circuit, but it is really an electric field, whether there happen to be any circuits or conductors about or not. Putting all this concisely: moving electric field causes magnetic field; moving magnetic field causes electric field.

Without going into the rather advanced mathematics of the thing, one might guess the possibility of the two fields in motion keeping one another going. Such a guess would be a good one, for they can and do. The speed they have to keep up in order to be mutually sustaining is the same as when they are guided by the transmission line. The combination of fast-moving fields is known as an *electromagnetic wave.* The union of the two fields is so complete that each depends entirely on the other and would disappear if it were destroyed. The speed of the wave is the same as that of light—a significant fact, which led to the discovery that light waves are the same as those used for radio communication, except for their much higher frequency.

17.5 Radiation

An important thing to note about the rope trick is that the rope itself does not travel away from us—only a wavy pattern. But there is another thing: some of the energy we put into it travels to the far end, for it is capable of setting a support there into vibration. Electromagnetic waves are similar in this respect too; they convey energy away from the source. The outward movement of the waves from the source is called *radiation.* All variations in current cause radiation, but when the field energy is mainly localised, say between the conductors of a transmission line or the plates of a capacitor, the amount is much less than when it is spread out as in Fig. 17.4. Radiation also increases with the rate of variation; the 50 Hz variations of ordinary a.c. power systems cause negligible radiation, but 50 MHz is quite a different matter. So although we have spoken of the waveforms of low-frequency a.c., this should not be taken as suggesting that they are waves in the sense used in this chapter—electromagnetic waves.

We saw in connection with the lumped tuned circuit of Fig. 14.2 that the electric and magnetic fields, responsible for the capacitance and inductance, are quarter of a cycle out of phase with each other. Even in the most spread-out circuit, such as our half-wavelength rod, the majority of the energy of the surrounding fields is able to return to the circuit twice each cycle, to be manifested there as its distributed capacitance and inductance. The majority portion of each field, mainly near the source, is similarly out of phase with the other. it is distinguished by the term *induction field,* and it falls off inversely as the square of the distance, which means that most of it is quite close to the source.

In order to convey energy right away from the source, however, there have to be components of fields in phase with one another, and it is this portion, called the *radiation field,* in which we are interested now. It falls off inversely as the distance, which means that although it starts as a minority it spreads much more, and far away from the source it is very much in the majority.

17.6 Polarisation

Fig. 17.4 shows that the electric and magnetic fields are at right angles to one another and to the direction in which they radiate outwards as electromagnetic waves. Imagine we are looking towards the vertical source and the oncoming waves. Then the electric field alternates up and down and the magnetic field side to side, as shown in Fig. 17.6. This pattern can be imagined as due to a transmission line consisting of parallel strips top and bottom. At the instant depicted, the potential is positive at the top, so the electric field is downwards from it. This p.d. would be driving current towards us through the top conductor and away from us through the bottom one. The 'corkscrew rule' (Fig. 4.2) indicates a clockwise magnetic field round the bottom current and anticlockwise round the top one; both combine to give a magnetic field from left to right, as shown.

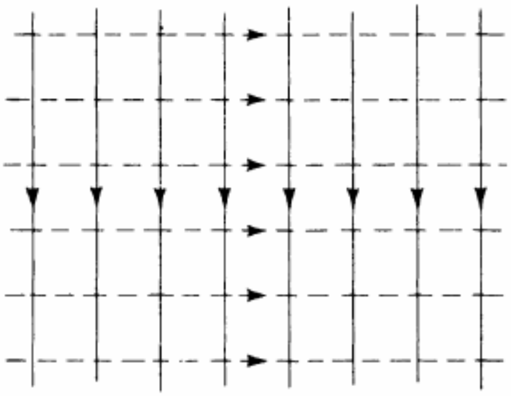


Fig. 17.6 Relative direction of electric (full line) and magnetic (broken line) fields in an electromagnetic wave advancing towards the viewer

If for any reason we prefer the directions of the two fields to be interchanged—electric horizontal and magnetic vertical—that can easily be arranged, by turning the radiator horizontal. Similarly for any other angle. The direction of the field in an electromagnetic wave is known as its *polarisation.* Since there are two fields, mutually at right angles, there would be ambiguity if it had not been agreed to specify the polarisation of a wave as that of its *electric* field. This custom has the advantage that the angle of polarisation is the same as that of the radiator.

Because radiation starts off with a certain polarisation it does not follow that it will arrive with the same. If it has travelled far, and especially if on its way it has been reflected from earth or sky, it is almost certain to have become disarranged. Another thing is that polarisation is not necessarily confined to a single angle. After reflection, radiation often has both vertical and horizontal components, and the maximum may be at some odd angle. If it is equal at all angles, the wave is said to be circularly polarised, because at any point it consists of fields of constant strength rotating continuously at the rate of one revolution per cycle. Sometimes for special purposes the radiator is designed to make the waves circularly polarised from the start.

17.7 Antennas

Having considered the process of radiation, we are due to turn attention to the radiator. If intended for that purpose it is called an aerial or *antenna.* The form we developed from a quarter-wavelength transmission line, known as a half-wave *dipole,* is perhaps the most important both from the theoretical and (since the arrival of f.m. broadcasting and television) practical point of view.

We have already noted that this straight metal wire or rod is a tuned circuit, and that its sound-wave analogue is a stretched cord as in Fig. 1.3. The fundamental frequency at which it resonates is the one that makes half a wavelength equal to (or, to be precise, a few per cent greater than) the length of the rod.

There are certain higher frequencies at which the same length of antenna can be set into vibration at harmonic frequencies; but because of complications due to currents flowing in opposite directions, tending to cancel out radiation, this condition is seldom used. It is not essential for an antenna to be in resonance at all, but as the radiation is proportional to the current, which is a maximum at resonance, antennas are practically always worked in that condition.

Just as with the tuned circuits in Chapter 8, the originator of the current has been shown as a generator in series; actually at the centre of the dipole. But in practice a tuned circuit is often 'excited' by an external field, as for example a magnetic field in inductive coupling. The same applies to our metal rod. If it is exposed to electromagnetic waves, the magnetic field pattern sweeping across it induces an e.m.f. in it. This statement must not be taken to mean that if there were no metal there would be no e.m.f. It is the movement of the magnetic field that induces the electric field, as we have seen; and that electric field constitutes an alternating e.m.f. in space.

It is unimportant whether the e.m.f. acting on the antenna is credited to the electric field or the magnetic field—they are simply two different descriptions of a single action—but it is usually convenient to work in terms of the electric component of the waves. Near a radiator it is reckoned in volts per metre, or microvolts per metre if far off. If a vertical wire two metres long is exposed to vertically polarised radiation of $100\,\mu V$ per metre, then an e.m.f. of $200\,\mu V$ will be induced in it. But if either the wire or the radiation is horizontal, no e.m.f. is induced. In general, e.m.f. is proportional to the cosine of the angle between the wire and the direction of polarisation.

It is in this way that an antenna is used at the receiving end. The only difference in principle between a sending and a receiving antenna is that one is excited by a r.f. generator connected to it, and the other by the electromagnetic waves created by the former. The factors that make for most effective radiation in any particular direction are identical with those for the strongest reception from that direction, so whatever is said about antennas as radiators can readily be adapted from the reception standpoint.

17.8 Radiation Resistance

We have noted that the dipole antenna, like the quarter-wave line from which we developed it, has inductance and capacitance. What about resistance? There is of course the resistance of the metal of which it is made, together with the various sources of loss resistance described in Secs. 8.14 and 15. But we have seen that the waves radiated from a sending antenna carry away energy from it, otherwise they would not be able to make currents flow in receiving antennas. It is convenient to express any departure of energy from a circuit in terms of resistance (Sec. 2.18). But whereas the other kinds of antenna resistance are wasteful, this kind, called *radiation resistance,* measures the ability of an antenna to do its job.

The radiation resistance of an ordinary tuning circuit is generally very small compared with its other resistance (radiation, after all, is not its job!); at the other extreme the loss resistance of a dipole, made perhaps of copper tubing, is small compared with its radiation resistance, which at resonance is about 70 Ω if it is measured at the centre where the current is greatest. If therefore a current of, say, 2A can be generated there, the power radiated, calculated in the usual way as I^2R, is $2^2 \times 70 = 280$ W. If the radiation resistance were measured at some other point along the dipole it would be greater, but this would not mean greater radiation, because the current there would be less.

A receiving antenna too has radiation resistance, and some of the energy imparted to it by the waves is re-radiated—a fact of great importance in the design of receiving antennas, as we shall see.

17.9 Directional Characteristics

We based our study of radiation from a dipole on the production of waves along a rope. But of course there is no reason for electromagnetic radiation to be confined along a single line like this; it goes out equally in all directions around the 'equator'. So the wave pattern in this plane can be likened to the ripples created on the surface of a pond by dropping in a stone. A skeleton of it could be produced by having ropes radiating from one's hand like the spokes of a horizontal wheel. But what about other angles, upwards and downwards?

directions, At *b* is the vertical polar diagram, showing how the radiation varies from maximum at right angles to the dipole to nil end-on. So if a dipole is placed horizontally it can be used to cut out undesired reception from two opposite directions. As regards reception, these diagrams assume, of course, that the radiation arriving has the correct polarisation. A horizontal dipole does not respond to vertically polarised waves from any direction.

17.10 Reflectors and Directors

Any metal rod is a tuned circuit which will have maximum currents set going in it by waves twice its own length. These currents radiate waves, just as if they were caused by a directly connected generator. If such a rod is placed close to a radiating dipole, the re-radiation from it is strong enough to modify the primary radiation very considerably, subtracting from it in some directions and adding to it in others. When the parasitic antenna (as it is called) is about the same length as the driven one, and quarter of a wavelength away from it, the phase of the re-radiation is such as to reverse much of the radiation beyond the rod, which is therefore called a *reflector*—see Fig. 17.9. The same principle applies to a receiving antenna, and reflectors are frequently used; not so much to strengthen the wanted signal (though that is usually helpful) but to cut out interference from the opposite direction.

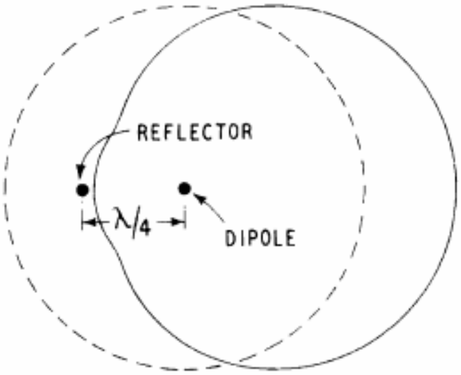


Fig. 17.9 Showing how the equal-all-round radiation of Fig. 17.8 (*a*) (repeated here dashed) is modified by an unconnected dipole placed quarter of a wavelength away from the first, and of such a length as to have a 45° phase angle

The effect just described depends on the phase relationship between the primary and secondary radiation. When considering tuned circuits we may have noticed (from Fig. 8.8, for example) that a small departure from exact tuning causes a large phase change between voltage and current. It is the same with dipoles; a small increase or decrease in length alters the phase and therefore the polar diagrams. The phase can also be varied by the spacing between the rods. If the parasitic rod is shorter than the driven one, and only about 0.1 λ away from it, the effect is opposite to that shown in Fig. 17.9, so such a rod is called a *director*. The use of reflectors or

Surprisingly a vertical slot radiates horizontally-polarised waves. In fact the slot can be regarded as a dual of the equivalent dipole. The dipole is a λ/2 conductor in non-conducting space; the slot is a λ/2 non-conductor in a conducting sheet. The electric field from the slot corresponds with the magnetic field of the dipole; the magnetic field from the slot corresponds with the electric field of the dipole.

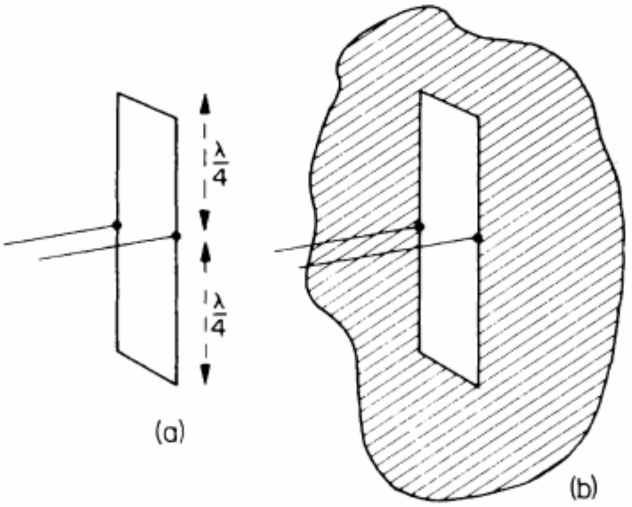


Fig. 17.11 Two λ/4 short-circuited lengths of twin-wire transmission line (*a*) become a slot antenna when conductor width is indefinitely extended (*b*)

Slot antennas are used at v.h.f./f.m. senders. The mast top is a vertical conducting cylinder with vertical slot radiators at 90° intervals around it to give unidirectional radiations.

17.11 Antenna Gain

Concentrating the radiation in the direction or directions where it will be most useful increases the apparent power of the sender. The number of times the intensity of radiated power in any particular direction is greater than it would be if the same amount of power were radiated equally in all directions is called the *power gain* (or just 'gain') of the antenna in that direction. For example, if the radiated power were concentrated uniformly into one tenth of the whole surrounding sphere the gain would be 10. Within that beam, the radiation would be the same as if ten times the power were being radiated equally in all directions. The effective power of a sender is therefore equal to the actual radiated power multiplied by the antenna gain; this figure is known as the *effective radiated power* (e.r.p.).

In broadcasting to the area around the sender, the aim is to increase the e.r.p. by directing as much of the radiation as possible horizontally, rather than upwards where it would be wasted. The apparent power of a sender, so far as any particular receiver is concerned, can of course, be further increased by using a receiving antenna with gain in the direction of the sender.

A simple dipole, by favouring the equatorial as compared with polar directions, has a gain of 1.64, and the gains of v.h.f. and u.h.f. receiving antennas are sometimes quoted relative to this rather than to the theoretical spherical radiator.

17.12 Choice of Frequency

A feature of the dipole antenna and its derivations which we have considered so far is that their dimensions are strictly related to the wavelength and therefore to the frequency (Sec. 1.9). What decides the choice of frequency: a convenient aerial length, or something else?

The shorter the wavelength and higher the frequency, the smaller and cheaper the antenna and the more practical it is to direct its radiation and increase its gain. For example, as long ago as 1931 a radio link was established across the English Channel on a wavelength of 17 cm, a half-wave antenna for which is only about 8.89cm long! Why, then, erect at vast expense antennas hundreds of metres high?

The main reason is that such waves behave very much like the still shorter waves of light, in not being able to reach anywhere beyond

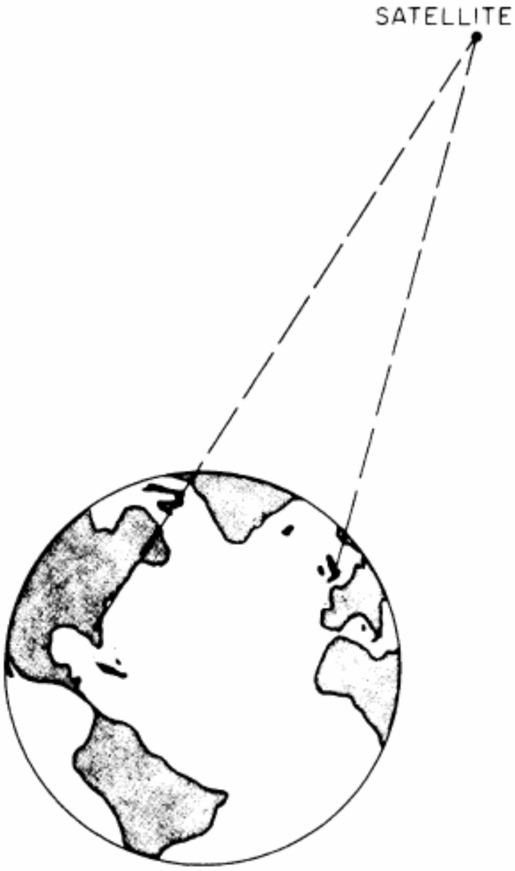


Fig. 17.12 'Geostationary' satellite for long-distance microwave communication

the direct line of 'view' from the sending antenna. So if the cheapness is to be maintained the range is usually restricted to a few kilometres. It can be increased to one or even two hundred kilometres by setting up the antennas on high towers or mountains, and to thousands of kilometres by using man-made satellites, for re-transmission, as in Fig. 17.12. The most useful type of satellite for this purpose is in orbit at such a distance (35 860 kilometres) that its period of revolution around the earth is 24 hours, so that if established in the same direction as the earth's daily rotation the satellite maintains a constant station relative to places on earth. It is therefore suitable for both television broadcasting and point-to-point communication.

Wavelengths shorter than 10 m are used for television, not because the shortness of wavelength in itself has any particular merit for that purpose, but because it corresponds to a very high frequency, which, as we saw in Chapter 15, is necessary for a carrier wave that has to carry modulation frequencies of up to several MHz. They are also used for short-distance communication such as police cars, radiotelephone links between islands and mainland, and other specialised short-range purposes.

17.13 Influence of the Atmosphere

At various heights between 96 and 480 kilometres above the earth, where the atmosphere is very rarified, there exist conditions which cause radio waves to bend, rather as light waves bend when they pass from air into water. The very short (high-frequency) waves we have just been considering normally pass right through these layers and off into space, as shown in Fig. 17.13*a*. They are therefore suitable for communication with and control of spacecraft. Where they are close to the ground, they are rapidly absorbed, as indicated by the shortness of the ground-level arrow. So their effective range there is limited.

As the frequency is reduced below about 30 MHz—the exact dividing line depends on time of day and year, sun-spot activity, and other conditions—the sky wave is bent so much that it returns to earth at a very great distance, generally several thousands of kilometres. Meanwhile the range of the wave-front travelling along the surface of the earth (and therefore termed the ground wave) increases slowly (Fig. 17.13*b*). Between the maximum range of the ground wave and the minimum range of the reflected wave there is an extensive gap, called the skip distance, which no appreciable radiation reaches.

As the frequency is further reduced this gap narrows, and earth reflections may cause the journey to be done in several hops (*c*). Since the distances at which the sky waves return to earth vary according to time and other conditions as mentioned, it is rather a mercy of various effects that make long-distance short-wave communication unreliable. For this reason they were originally selected as the only wavelengths suitable for long ranges, but now only a small fraction of long distance communication is borne by very long waves. The disadvantages of the latter are (1) the enormous size of antenna needed to radiate them; (2) the low efficiency of radiation even with a large and costly antenna system; (3) the higher power needed to cover long ranges, due largely to (4) the great intensity of atmospherics—interference caused by thunderstorms and other atmospheric electrical phenomena; and (5) the very limited number of stations that can work without interfering with one another, because the waveband is so narrow in terms of frequency—which is what matters; see Chapter 15.

17.14 Earthed Antennas

We see, then, that although for convenience of dipole antenna size one would choose frequencies of a few hundred MHz, these are not suitable for all purposes. For distances of more than about 80 kilometres it is usually necessary to use longer waves (lower frequencies). Until now we have assumed that the antenna is suspended well clear of the earth, but this is not convenient with a dipole for wavelengths of many metres, and the length of the antenna itself may be unwieldy. Marconi used the ground itself as the lower half of a vertical dipole; the remaining half sticking up out of it is therefore often called a quarter-wave Marconi antenna. The current and voltage distribution are shown in Fig. 17.14, which is the same as the top half of Fig. 17.2.

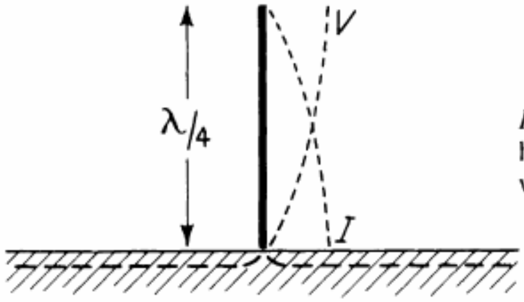


Fig. 17.14 The earth can be used as one half of a radiator, leaving a quarter-wavelength wire above earth

For the earth to be a perfect substitute for the lower half of a dipole it must be a perfect conductor, which of course it never is. Sea water is a better approximation. To overcome the loss due to imperfectly conducting earth, it is a common practice to connect the lower end of the antenna to a radial system of copper wires. An insulated set of wires stretched just above the ground is known as a *counterpoise*. Alternatively, to avoid tripping up people or animals, the wires are often buried just below the surface, as shown dashed in Fig. 17.14.

17.15 Feeding the Antenna

One advantage of the Marconi antenna is that it places the middle of the dipole at ground level, which is helpful for coupling it to the sender or receiver. Even so, it is not always convenient to have the output circuit of a sender close enough to the foot of the antenna to couple it directly, and it is seldom convenient to have either sender or receiver at the centre of a suspended dipole, so there is a need for some means of linking the antenna with the ground equipment. Chapter 16 showed how the problem can be conveniently and efficiently solved by means of a transmission line designed to match the impedance of the antenna. The object of matching, we recall, is to avoid reflections, which prevent the maximum transfer of power to or from the antenna and cause excessive line currents and voltages at a sender, or signal echo effects at a receiver.

We have noted that an ordinary centre-fed dipole has a radiation resistance of about 70 Ω. At resonance, which is the normal working condition, inductance and capacitance cancel one another out. Other kinds of resistance are usually negligible, so the dipole can be treated as a 70 Ω load. The simplest scheme is to use a line having a characteristic resistance, R_0, of about 70 Ω, as in Fig. 17.15. Here the dipole can be regarded as a radiating extension of the line. At the other end, the line should be matched to the sender output circuit.

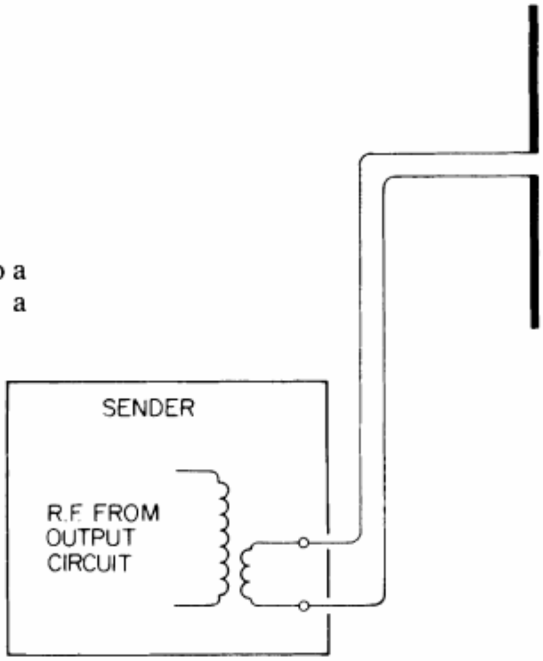


Fig. 17.15 A simple dipole connected to a sender (or receiver) by means of a matched-impedance transmission line

Fig. 16.4 shows that 70 Ω is an awkward R_0 for a parallel-wire line, for the wires have to be almost touching; it is also a long way from the lowest-loss value. On the other hand it is about optimum for coaxial cable. Moreover the coaxial type radiates less and picks up less interference. But it has the disadvantage of being into one half dipole and the other half into the other. Next, suppose the energising of the two half antennas is entrusted to separate generators (*c*) each supplying $I/2$. If they are in phase with one another, the ends of the two dipoles will be at the same potential so can be joined together without making any difference. Finally (*d*) the two voltages V are supplied by one generator, which must therefore provide twice the voltage of the original one (*a*) to make half the original current flow into the modified antenna, called a *folded dipole.* Its resistance is therefore four times that of an ordinary dipole, so is matched by a 280 Ω feeder.

A folded dipole is shown as the main element in the Yagi antenna in Fig. 17.10, because with many directors and reflectors the impedance of a dipole is much reduced and a folded type may be a better match than an ordinary one for a coaxial line. Moreover, the centre point of the non-fed half is at zero potential so needs no insulation if the antenna is supported there.

17.16 Tuning

A dipole antenna is self-tuned, by virtue of its chosen length, to a particular frequency. But an antenna is often required to work over a very wide range of frequency, especially for receiving. Instead of varying the length of the antenna wire, it is usual to augment its distributed inductance and capacitance by means of variable capacitors or inductors, as for example in Fig. 17.17. These tuning components reduce the resonant length of the antenna for a given frequency, which may be very desirable; for example, even a quarter-wave self-tuned antenna for 30 kHz would have to be over 12000 metres high! It is quite usual for a considerable proportion of the reactance of an antenna system, especially for medium and long waves, to be in concentrated form; but it must be understood that this is at the expense of radiation or reception.

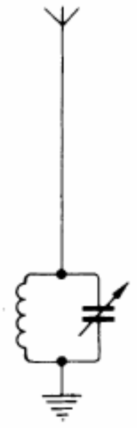


Fig. 17.17 A common method of tuning an antenna to a desired wavelength

Where height is limited, by restrictions imposed by aviation or by the resources of the owner, it is possible to increase radiation moderately by adding a horizontal portion, as in the T or inverted L antennas, Fig. 17.18. The effect of the horizontal extension is to localise the bulk of the capacitance in itself so that the current in the vertical part, instead of tailing off to zero, remains at nearly the maximum and therefore radiates more. The top portion radiates too, but with a different polarisation, so the addition due to it may not be very noticeable in a vertical receiving antenna.

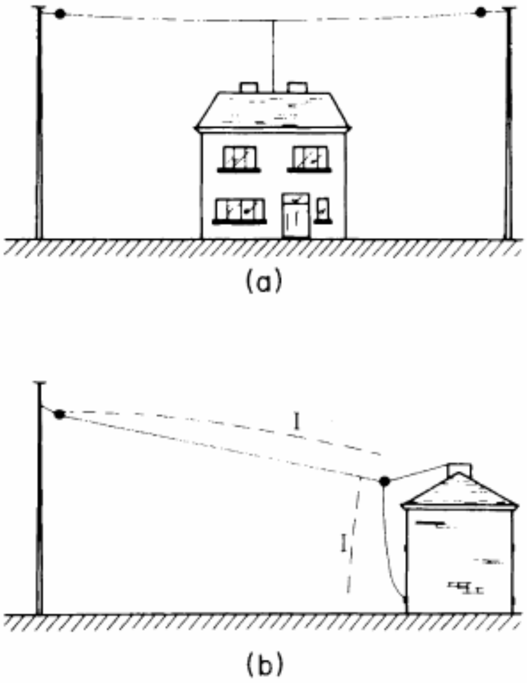


Fig. 17.18 To increase the radiation from a vertical antenna without increasing its height, a horizontal top is commonly added, forming (*a*) the T or (*b*) the inverted L type. Note the more uniform current distribution in the vertical part (*b*)

17.17 Effective Height

In general, the greater the radiation resistance of an antenna the more effective it will be for sending. But a more useful piece of information about an antenna when it is used for receiving is what is called its *effective height.* It is the height which, when multiplied by the electric field strength, gives the e.m.f. generated in it. For example, if the wave from a sender has a strength of 2 mV per metre at a receiving station, and the effective height of the antenna is 5 metres, the signal voltage generated in the antenna will be 10 mV. The effective height is less than the physical height. It can be calculated for a few simple shapes, such as a vertical wire rising from flat open ground, but usually has to be measured, or roughly estimated by experience.

17.18 Microwave Antennas

What happens if, instead of making the wavelength longer than is most handy for dipoles, it is made much shorter—say a few centimetres? The radiation from a single dipole is of course limited by its size, and to bring the e.r.p. (Sec. 17.11) up to a useful figure a relatively large reflector is needed. Instead of an enormous number of small dipoles, it is made of continuous metal, or, to reduce weight and wind resistance, wire netting. The currents induced in this by the driven dipole cause reflection, but in a more controllable pattern, for the metal can be formed into a parabolic reflector, like a large headlight or bowl heater, usually called a dish. In this way the radiation or reception can be confined very closely in the required direction.

Just as at these frequencies the two-conductor transmission line gives place to a waveguide (Sec. 16.10), the dipole gives place to a horn, which is an expanding end to the waveguide, arranged to direct the radio power on to the reflector.

Such frequencies and antennas are commonly used for radar, where narrow beaming is essential, and for point-to-point transmission of television signals.

17.19 Inductor Antennas

In all the antennas considered, the circuit between the terminals is completed by the capacitance from one arm to the other or to earth. Although they have inductance, they can be regarded roughly as opened-out capacitors. It is possible instead to use an opened-out inductor, in which of course there is a wire path all the way. This type is known as a frame or loop antenna, and because its loss resistance is usually larger than the radiation resistance it is seldom used as a radiator, but mainly for receiving, especially for direction finding.

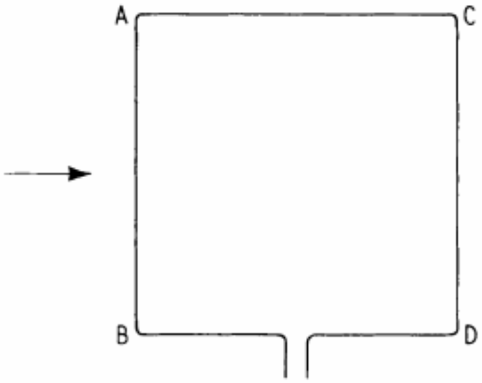


Fig. 17.19 The effectiveness of a receiving antenna consisting of a loop or coil of wire depends on the direction from which the waves are arriving

We noted that reception by an open antenna is usually calculated from the strength of the electric field, although the same answer would be given by the magnetic field. The opposite applies to loop antennas. Taking as example a single square turn of wire as in Fig. 17.19, one can see that waves arriving broadside on, whatever their

CHAPTER 18

Detection

18.1 The Need for Detection

Turning to the reception end in Fig. 1.11, we see a number of boxes representing sections of a receiver. Although all of these sections are included in most receivers, they are not all absolutely essential. So we begin with the detector, because it is the one thing indispensable, for reasons now to be explained.

At the receiving antenna, the modulated carrier wave generates e.m.fs whose waveforms exactly match those of the currents in the sending antenna. The frequency of those currents, as we know, is so high as to be inaudible. It has been chosen with a view to covering the desired range most economically (as discussed in the previous chapter), and bears little relationship to the frequencies of the sounds being transmitted. In order to reproduce these sounds, it is necessary to produce currents having the frequencies and waveforms of the sounds. For example, given the amplitude-modulated carrier wave shown in Fig. 15.1*c* (or 18.1*a*), the problem is to obtain from it Fig. 15.1*a*. The device for doing so is the *detector*. It is an example of a type of circuit which extracts the modulation frequency from a modulated carrier as the desired product, and now commonly known as a demodulator. (This term was formerly used for an effect that removed interfering modulation as an undesired audible product; see Sec. 22.10. This is a logical use of the word; cf 'decontamination').

For receiving frequency-modulated signals (Sec. 15.4) it is necessary also to include in the demodulator means for converting the frequency variations into the amplitude variations on which the detector works. This device—the *discriminator*—is considered later on in this chapter.

18.2 The Detector

The simplest method of detection is to eliminate half of every cycle, giving the result shown in Fig. 18.1*b*. At first glance, this might not seem to differ fundamentally from *a*, for the original frequency appears to be still there, though at half-strength. It is quite true that the original frequency is there, and means have to be provided for getting rid of it. The fundamental difference is that whereas the average value of each cycle in *a* is zero (because the negative half cancels the positive half), in *b* it is proportional to the amplitude (at that moment) of the modulated carrier wave, which in turn is proportional to the instantaneous value of the a.f. modulating signal. The dashed line in Fig. 18.1*b* indicates the average level of the current, and it can be seen to vary at the same frequency and with the same waveform as Fig. 15.1*a*. There is admittedly some loss in amplitude, but we shall see that the types of detector in common use manage to avoid most of this loss and give an output almost equal to the peak values.

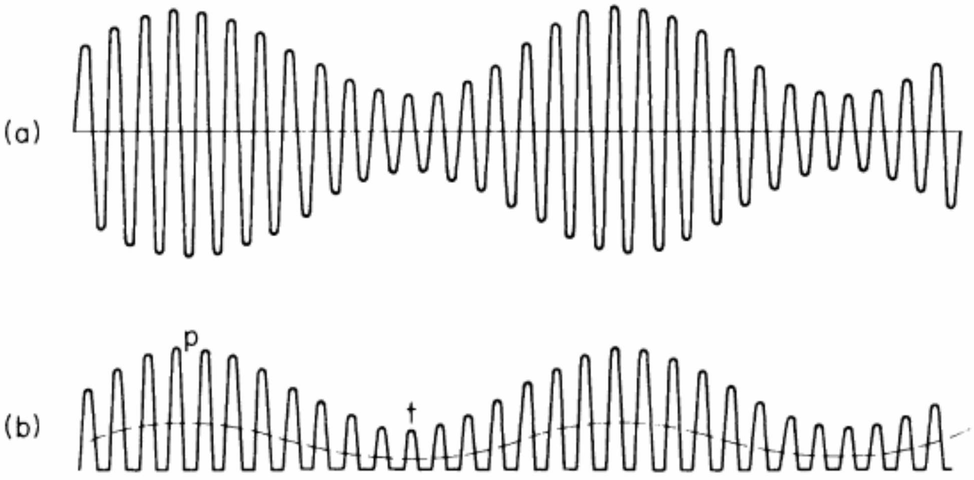
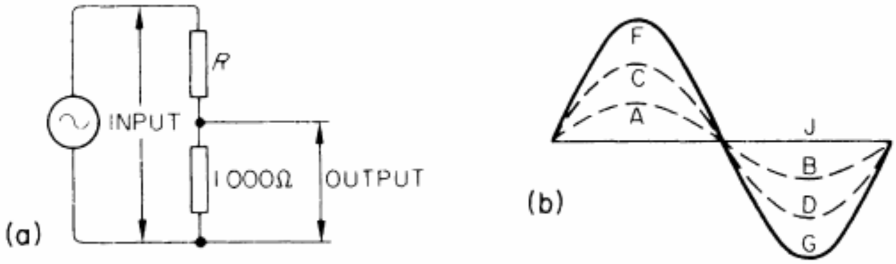


Fig. 18.1 (*a*) Sample of an amplitude-modulated carrier wave, similar to Fig. 15.1*c*. The average value of each cycle is zero: but if one half of each is eliminated (*b*) the average value of the other halves, shown dashed, fluctuates in the same manner as the modulating waveform (e.g. Fig. 15.1*a*)

The process of suppressing half of each cycle, converting alternations of current into a series of pulses of unidirectional current, is known by the general term *rectification* already met in Sec. 5.5. The detector usually consists of a rectifier adapted to the particular purpose now in view, and may be taken to include means for extracting the desired a.f. from a mixture like Fig. 18.1*b*. It should be noted that this rectified signal contains not only r.f. and a.f. components, but also a unidirectional component (d.c.). That is, the desired a.f. current does not alternate about zero, as in Fig. 15.1*a*, but about some steady current—positive or negative, depending on which set of half-cycles was eliminated by the rectifier. Although this d.c. is unnecesary and in fact undesirable for reproducing the original sounds, it has its uses as a by-product (Sec. 22.11).

18.3 Rectifiers

A perfect rectifier would be one that had no resistance to current flowing in one direction, and an infinite resistance to current in the resistance, and HOJ infinite resistance; so FOJ is the graph of a perfect rectifier. COB is an example of a partial rectifier—a resistance of 500 Ω to positive voltages and 2000 Ω to negative voltages. Now apply an alternating voltage to a circuit consisting of any of the foregoing imaginary resistances in series with a fixed resistance of 1000 Ω (Fig. 18.3*a*). The voltage across the 1000 Ω—call it the output voltage—is equal to the applied voltage only if *R* is zero. Both voltages are then indicated by the full line in Fig. 18.3*b*. If *R* were 500 Ω (COD), the output would be represented by the dashed line CD, and if 2000 Ω by the dashed line AB.



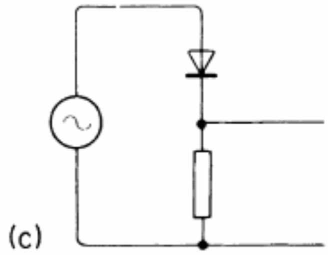


Fig. 18.3 The results of applying a sinusoidal voltage to a circuit (*a*) consisting of any of the resistances represented by Fig. 18.2 (denoted by R) in series with 1000 Ω are indicated by (*b*), which is lettered to correspond with the various characteristics in Fig. 18.2. When R is a rectifier it is represented as in (*c*)

If now a rectifier is substituted for *R*, as shown in Fig. 18.3*c*, the positive and negative output voltages are unequal. For example, the perfect rectifier (FOJ) gives the output FJ in *b*; and the partial rectifier gives CB. The average voltage, taken over each whole cycle, is, of course, nil when *R* is not a rectifier. Using the perfect rectifier, the average of the positive half is 63.6% of the maximum (Sec. 5.4), while the negative is nil; so the average rectified current taken over a whole cycle is 31.8% of the peak value. Using the rectifier with the graph COB, the negative half-cycle (B in Fig. 16.3*b*) cancels out half of the positive half (C), which is two-thirds of the input; so the rectified result is less than 11% of the peak input.

18.4 Linearity of Rectification

Whether complete or partial, however, the degree of rectification from rectifiers having characteristics like those in Fig. 18.2 would be independent of the amount of voltage, provided, of course, that the rectifier was arranged to work exactly at the sharp corner in its characteristic, whether that happened to correspond with zero voltage, as in Fig. 18.2, or not. So if we were to draw a graph showing the net or average or rectified voltage output against peak alternating voltage input to a rectifier of this kind we would get a straight line, such as the two examples in Fig. 18.4, corresponding to characteristics FOJ and COB in Fig. 18.2. These are therefore called linear rectifiers, it being understood that the term is in reference to a graph of this kind and not of the Fig. 18.2 kind.

It should be clear from Fig. 18.1*b* that it is necessary for the rectifier to be linear if the waveform of the modulation is not to be distorted. It is therefore unfortunate that in reality there is no such thing as a perfectly linear rectifier. The best one can do is to approach very close to perfection by intelligent use of the rectifiers available. The weak feature of all of them is the more or less gradual transition from high to low resistance, exemplified by the typical 'bottom bend' in Fig. 9.3, which causes weak signals to be rectified less completely than strong ones. Such a characteristic is represented in Fig. 18.4 by the curved line, and is explained in greater detail later (Sec. 18.11). Semiconductor diode characteristics are considerably better, if we allow for the offset or displacement from zero, especially with silicon (Sec. 9.11).

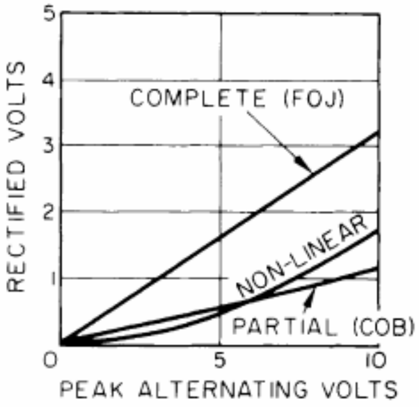


Fig. 18.4 The rectified outputs obtained by applying alternating voltages to various rectifiers are shown in these characteristic curves, lettered to correspond with Figs. 18.2 and 18.3

18.5 Rectifier Resistance

There were two reasons for the poor results of rectifier COB in Fig. 18.2 compared with FOJ. One was its *forward resistance* (500 Ω), which absorbed one-third of the voltage, leaving only two-thirds for the 1000 Ω load. The other was its finite *backward resistance* (2000 Ω) which allowed current to pass during the negative half-cycles, so neutralising another one-third of the voltage. With such a rectifier it would clearly be important to choose the load resistance carefully in order to lose as little as possible in these ways. As it happens, 1000 Ω does give the largest output voltage in this case, the general formula being $\sqrt{(R_F R_B)}$, where R_F and R_B are respecitvely the forward and backward resistances of the rectifier.

As compared with crystal diodes, the vacuum diode does have one point in its favour. Although it is not perfect in the forward direction, for it always has some resistance, it is practically perfect in the reverse direction; that is to say, it passes negligible backward current, even when quite large voltages are applied. So even if, owing to a high forward resistance, the forward current were very small, the output voltage could, theoretically at least, be made to approach that given by a perfect rectifier, merely by choosing a sufficiently high load resistance. Silicon diodes are now available with very high backward resistance, and lower forward resistance than thermionic types. They have a sharper rectification bend, and of course need no cathode heater. But they need a little forward bias.

Although there are limits to the load resistance that can be used in practice, the loss of rectification can usually be made insignificantly small. In fact, by a very simple elaboration of Fig. 18.3*c*—a capacitor in parallel with the load resistance—it is possible to approach a rectification efficiency of 100% instead of the 32% that is the maximum without it, and so do nearly three times better than our 'perfect' rectifier. This point deserves closer consideration.

18.6 Action of Reservoir Capacitor

The modified rectifier circuit is now basically as Fig. 18.5. Suppose, in order to get the utmost rectified voltage, we made the load resistance *R* infinitely great, and that the backward resistance of the diode was also infinite. The capacitor C would then be in series with the diode, with no resistor across it. To simplify consideration we shall apply a square wave instead of a sine wave; amplitude 10 V (Fig. 18.6a). We shall also assume that its frequency is 500 kHz, so that each *half* cycle occupies exactly one millionth of a second (1 μs). At the start C is uncharged, and therefore has no voltage across it. It can acquire a charge only by current flowing through the rectifier, which offers a certain amount of resistance, and so when the first positive half-cycle arrives its 10 V at first appears wholly across that resistance, as shown by the full line V_D in Fig. 18.6*b*. The current driven by this voltage starts charging the capacitor, whose voltage thereby starts to rise towards 10 V, as shown by the dashed line. It is

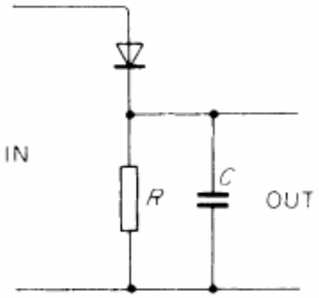


Fig. 18.5 Diode rectifier with the load resistor (*R*) shunted by a reservoir capacitor, *C*

Because C maintains the voltage during the half-cycles when the supply is cut off, it is called a *reservoir capacitor.*

18.7 Choice of Component Values

This, of course, is excellent so far as an unmodulated carrier is concerned. When it is modulated, however, the amplitude alternately increases and decreases. The increases build up the capacitor voltage still further; but the decreases are powerless to reduce it, for the capacitor has nothing to discharge through. To enable V_C to follow the modulation it is necessary to provide such a path, which may be in parallel with either the capacitor or the diode, so long as in the latter case there is a conducting path through the input to complete the discharge route. The resistance of the path should be low enough to discharge the capacitor C at least as fast as the modulation is causing the carrier amplitude to decline. The speed of decline increases with both depth and frequency of modulation, and calculating the required resistance properly is rather involved, but a rough idea can be obtained as follows.

Suppose, for example, the highest modulation frequency is 10 000 Hz. Then the time elapsing between the peak of modulation and trough (*p* to *t* in Fig. 18.1*b*) is half a cycle, or 1/20 000th of a second (50 μs). If C had to discharge at least 63% of the difference in carrier amplitude in that time, the time constant would have to be not more than 50 μs, and to avoid appreciable loss and distortion of top frequencies had better be less. Of the two factors making up the time constant *CR, R* is decided on the grounds just about to be considered, leaving us with the choice of *C* to make the time constant suitable.

We have no doubt realised by now that the discharge resistance *R* needed to enable the rectified output to follow the modulation is the same thing as what previously we regarded as the load resistance. If the load is the final outlet of the receiver, such as a pair of headphones, then its impedance is limited in practice to a few kilohms. More often however the detector is followed by an amplifier to increase the power to a level suitable for working a loudspeaker or television tube. With an f.e.t. amplifier the input impedance is many megohms—too high and indefinite to serve as the discharge path. So something of the order of 0.5 MΩ is chosen for *R*. This suits the thermionic diode but is on the high side for semiconductor diodes. The value of *C* that makes *CR* equal to 50×10^{-6} is 100 pF. For good reproduction of the highest audio frequencies either *C* or *R* or both should be reduced. The limit begins to be reached when *C* discharges appreciably between one r.f. peak and the next, causing a loss of detection efficiency at all modulation frequencies. The choice of *CR* is most difficult when the carrier-wave frequency is not many times greater than the top modulation frequency, and for this reason (among others) the carrier frequencies chosen for 'hi-fi' broadcasts are always very high (v.h.f.).

If a bipolar transistor is used as the first modulation amplifier, its input resistance is normally of the order of one or two kilohms, or rather more if it is increased by means of an unbypassed emitter biasing resistance (Sec. 12.9). To maintain a sufficiently high reservoir action at the lowest r.f. (which we shall see in Chapter 21 is usually 470 kHz or thereabouts) C must be much greater, usually about 10 nF, which with 4 kΩ gives a time constant of 40 µs.

The diode is chosen for a forward resistance that is small and a backward resistance large compared with R; for R equal to a few kilohms this requirement is readily met by point-contact germanium types.

18.8 The Diode Detector in Action

As the action of the deceptively simple-looking circuit in Fig. 18.5 is not very easy to grasp, and is also most important, it will be worth spending even more time on it. Fig. 18.7 is a diagram which, compared with Fig. 18.6, is speeded up so as to show some modulation, and uses proper sine waves for the carrier. Also, the voltage V_C across the capacitor—which, of course, is the output voltage—is drawn together with the r.f. input voltage V_I. Notice that whenever V_I is more positive than V_C the diode conducts and C charges rapidly, because the time constant is CR_F, R_F (the forward resistance of the diode) being much less than R. Directly V_I drops below C_V, C discharges through the only path available (R) with the much longer time contant CR. It is too long in this case, because V_C loses contact with V_I altogether during the troughs, and its waveform is obviously a distorted version of the modulation envelope. But that is because for the sake of clearness the carrier frequency shown is excessively low for the modulation frequency.

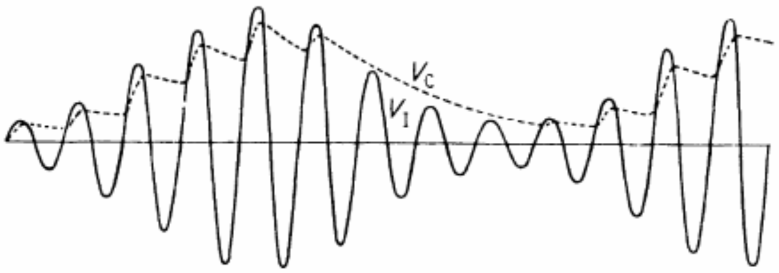


Fig. 18.7 V_1 represents a modulated r.f. input voltage and V_C the resulting output voltage

Note too that in addition to the desired frequency, V_C has also a saw-tooth ripple at the carrier frequency. Not only has C improved the rectification efficiency but it has also got rid of most of the unwanted r.f.

The third component making up V_C is a steady positive voltage (zero frequency), represented by the average height of the dotted line above the zero line.

18.9 The Detector as a Load

In Chapter 20 we shall see that the *Q* of receiver tuning circuits is very important. The detector receives its input from such a circuit, across which is is connected. Doing so is bound to reduce the *Q* of that circuit, and how much it does so can be calculated (Sec. 8.16) if we know the resistance of the detector. But do we?

With a perfect diode in series with a resistance *R*, as was assumed in Sec. 18.2 and shown in Fig. 18.8*a*, the resistance during one set of half-cycles is *R* and during the other set is infinite, or, with a practical diode, so much larger than *R* as to have negligible effect. We know (Sec. 14.5) that the energy-storing capability of a circuit having a high enough *Q* to be useful for tuning enables its waveform and amplitude to continue almost unchanged even though the circuit receives its maintaining energy in pulse form. The same applies to the removal of energy by resistance. So a resistance load that exists only during every alternate half-cycles is virtually the same as half the load all the time. The power, or rate of energy, delivered to a resistance *R* is V^2/R, *V* being the voltage developed across *R* (Sec. 2.17). So half the load due to *R* all the time is the same as the load due to 2*R* all the time.

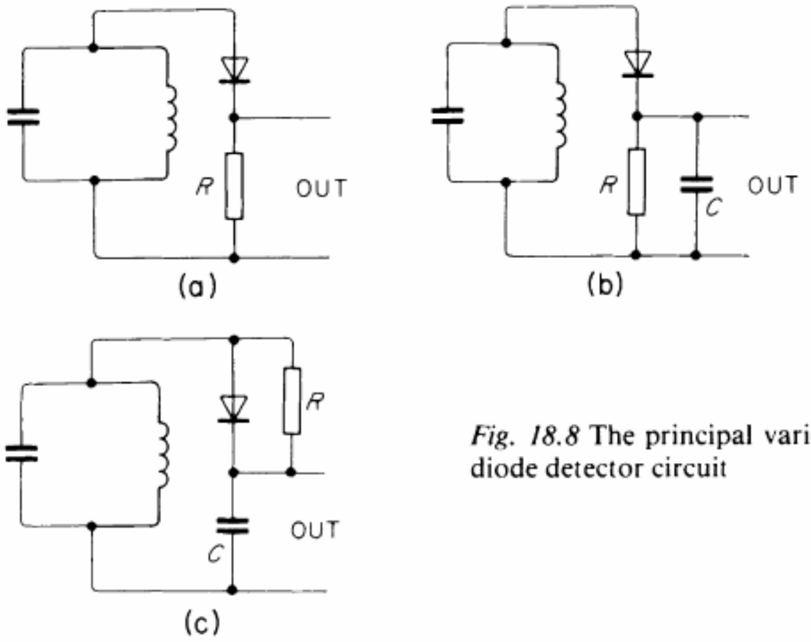


Fig. 18.8 The principal varieties of series diode detector circuit

What about the reservoir or peak detector, Fig. 18.8*b*? This looks difficult to assess because, as Fig. 18.7 shows, the tuned circuit is unloaded during most of each cycle, but for an unspecified period at each positive peak it is connected by the diode to *R* and *C* in parallel. But if we assume that the nearly steady voltage across *C* is equal to the peak input voltage (in practice, with suitable component values, it nearly is), we can get at it this way: Denote the resistance which, if connected across the tuned circuit in place of the detector, would have the same loading effect, by R'. If V is the r.m.s. voltage supplied by the tuned circuit, the output voltage according to the above assumption is $\sqrt{2}V$. As R is the only component dissipating power, that power will be $(\sqrt{2}V)^2/R = 2V^2/R$. But from the way in which we have defined R$'$ this power is also equal to V^2/R':

$$\frac{V^2}{R'} = \frac{2V^2}{R} \qquad \text{So} \quad R' = \frac{R}{2}$$

Provided that the input circuit is a low-resistance d.c. path, one end of R can be transferred to the other end of it as shown in Fig. 18.8*c*, but now (assuming C has negligible impedance to the carrier) R is across the r.f. input all the time, in addition to the detector loading as already calculated. So R' is now R and $R/2$ in parallel, which is equal to $R/3$, the loading effect being 50% more than in circuit *b*. This arrangement is therefore seldom used.

It should hardly be necessary to say that when an output of the opposite polarity is required the diode is simply reversed.

The above calculations of loading effects were all based on the assumption of a tuned input circuit. For some purposes (in measuring r.f. voltages, for example) a peak type of diode detector may be applied to a non-resonant circuit. This has no stored energy to supply the extremely non-uniform load, which requires strong pulses of charging current at each alternate peak of the input voltage. If the input source has much resistance these pulses cause a more or less considerable drop in voltage and therefore a loss of output voltage.

There is a drop in voltage even with a tuned circuit if its dynamic resistance is not small compared with the equivalent resistance of the detector. The few-kilohm load where the detector is followed by a transistor loads a tuned circuit excessively, so a suitable step-down transformer ratio from the tuned circuit is chosen.

18.10 Filters

The outputs from detector circuits *b* and *c* in Fig. 18.8 are something like the dashed line in Fig. 18.7. Both types of output contain the desired modulation signal, and both contain also a unidirectional

18.11 Detector Distortion

In discussing rectifiers we noted that they had to be non-linear conductors in order to rectify, but that if the non-linearity consisted of a sharp corner between two straight portions (e.g., COB in Fig. 18.2), the rectifier itself could be described as linear (e.g. COB in Fig. 18.4). And that no real rectifiers have perfectly sharp bends, so all are more or less non-linear. Now that we have seen the details of complete detector circuits we are returning to this topic, to consider how to ensure that the non-linearity is less rather than more.

The simple Fig. 18.8*a* rectifier circuit would be the easiest to analyse, but because the reservoir type is the one used almost invariably let us assume it, even though we shall have to do some 'guesstimating'. Fig. 18.12 shows the characteristic curve of a germanium diode of a type used for this purpose, and the load resistance (*R*) is say 10 kΩ. Let us suppose that an unmodulated r.f. signal of 1 V peak is applied. And let us guess that with this input the detector has an efficiency of 70%. In other words d.v. output across C and R which back-biases the diode is 0.7 V. So we draw a cycle of the input voltage from —0.7 V as its base line, between A and B. With this voltage across R, the continuous current through it must be 0.07

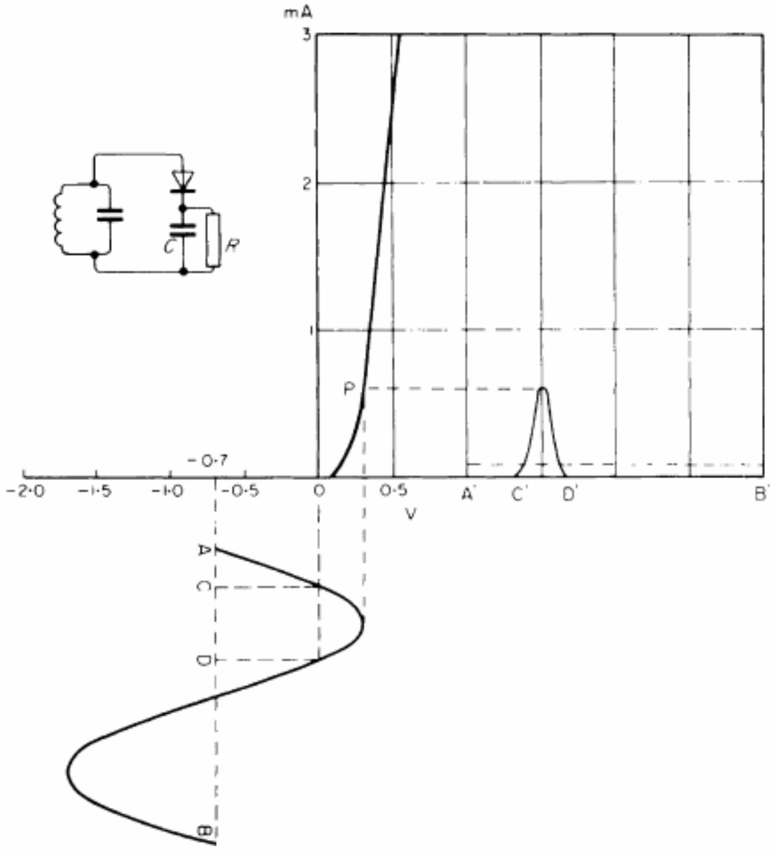


Fig. 18.12 Analysis of germanium diode detector performance by means of characteristic curve of the diode

In all the foregoing we have assumed that there is no capacitance across the diode. If there were it would form with C a potential divider across the signal source, reducing the efficiency of detection. For this reason junction diodes, which have enough self-capacitance to be appreciable, especially at v.h.f. and u.h.f., may be unsuitable; if so, point-contact types with their very small contact area are used. External stray capacitance must be avoided by attention to the circuit layout.

18.12 Shunt-diode Detector

Let us return to Figs. 18.5 and 18.7 for a moment. The voltage across the diode must, of course, be equal to V_1 minus V_C, and this is plotted in Fig. 18.13 together with Fig. 18.7 for reference. The signal across the diode, like that across the reservoir capacitor, has a.f. and z.f. components, and therefore could be used as the output voltage of the detector. The circuit diagram of such a detector is shown in Fig. 18.14. Because the diode is connected directly across the output, it is known as a shunt-diode detector. (The circuits described earlier, Fig. 18.8, are, of course, series-diode detectors.) The reactance of *C* at radio frequencies is small compared with the value of *R*, and nearly the full r.f. voltage from the *LC* circuit appears at the output. Thus the shunt-diode detector has a much greater r.f. ripple output than the series-diode circuits, making the use of an r.f. ripple filter essential. Moreover capacitor *C* effectively connects *R* across the resonant circuit as in Fig. 18.8*c* so that the loading effect on the tuned circuit is equivalent to a resistance of *R*/3. It is not surprising, therefore, that the series-diode circuit is normally preferred.

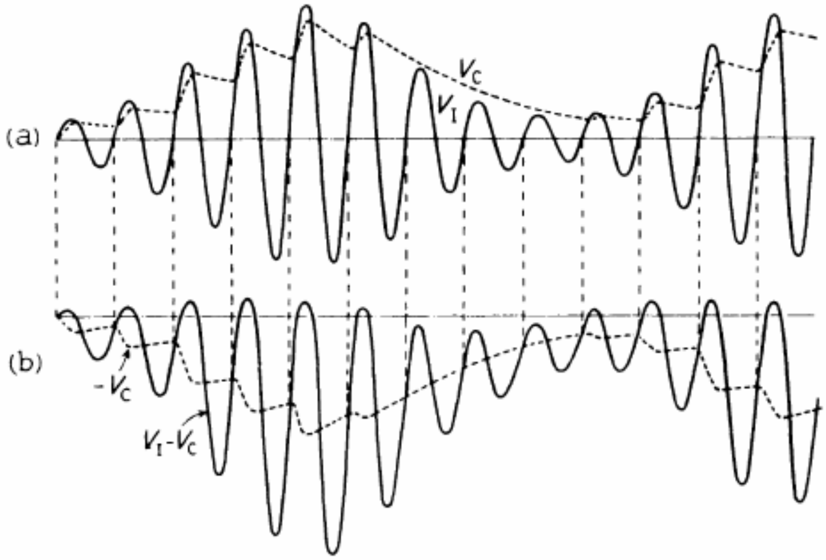


Fig. 18.13 In diagram (*a*), V_1 represents a modulated r.f. input voltage and V_C the resulting output voltage. In diagram (*b*), the full line shows V_1-V_C: i.e. the voltage across the diode

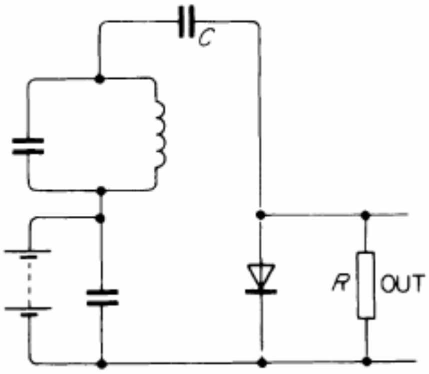


Fig. 18.14 Shunt diode detector circuit

However the shunt-diode circuit has one advantage. The reservoir capacitor *C* acts as a z.f. block between the tuned circuit and the diode, making the shunt-diode circuit useful where the *LC* circuit is at a d.v. potential which would upset the action of the diode if it reached it. This occurs, for example, when the tuned circuit is directly connected in the collector or anode circuit of an r.f. amplifier. The battery in Fig. 18.14 is intended to simulate these conditions.

18.13 Television Diode Detector

Some liberties can be taken in the value of the time constant for an a.m. sound detector because selectivity requirements (Sec. 15.9) in the receiver cause the r.f. circuits to restrict the pass-band and limit the upper audio frequencies to 4 kHz or so. Even if the receiver can accept the whole of the transmitted bandwidth, the upper modulation frequencies, being harmonics, are of low amplitude and give correspondingly shallow modulation.

In television, modulation depths up to 100 per cent are possible at the highest video frequencies; in fact the test card includes a grating which does just this. So in the UK system the detector must be capable of demodulating a 5 MHz video signal at full modulation depth. A repeat of the calculation made earlier, but substituting 5 MHz for 10 kHz, suggests that the time constant should be no greater than 20 ns. In fact higher values can be used because the resulting distortion of the waveform is imperceptible on the screen. A commonly-used value is 33 ns and, as the load resistance is made 3.3 kΩ, this gives a reservoir capacitance of 10 pF. Care must be taken to reduce to a minimum stray wiring and component capacitances which can easily amount to a few pF.

18.14 F.M. Detectors

The performance of all the detectors considered so far depends hardly at all on the frequency of the carrier wave between very wide limits. Consequently they treat a frequency-modulated carrier wave (Sec. 15.4) as if it were unmodulated. In order to receive f.m. signals it is necessary either to convert the variations of frequency into corresponding variations of amplitude so that an ordinary detector can be used, or else use a different kind of detector altogether.

The first method can be employed in a very simple manner with an ordinary a.m. receiver by adjusting the tuning so that instead of the incoming carrier-wave frequency coinciding with the peak of the resonance curve (Fig. 8.5) it comes in on one of the slopes. As its frequency rises and falls by modulation, the amplitude of response rises and falls. But because the slope of the resonance curve is not quite linear, the amplitude modulation is not a perfect copy of the frequency modulation, so there is some distortion. More serious still, the potential benefits of f.m. are not achieved, because the receiver is open to amplitude modulation caused by interference. And both selectivity and efficiency are poor.

Receivers designed for f.m. reception therefore include means for rejecting amplitude modulation and responding only to frequency modulation. The first of these functions is performed by a *limiter* and the second by a *discriminator*. In one of the most popular types of f.m. detector—the ratio detector—these functions are combined. However, for a first approach it will be easier to consider a system in which they are separate.

In connection with oscillators we saw in Sec. 14.9 that if the amplitude of oscillation exceeds a certain amount the transistor or valve cuts off the positive or negative peaks or both. A limiter is usually one of these devices so arranged that it does this at anything above quite a small signal amplitude, so that however much greater the incoming signal may be it is cut down or limited to the same size. Variations in amplitude are therefore almost completely removed, and with them much of the effects of interference.

18.15 Foster-Seeley Discriminator

Fig 18.15*a* shows the essentials of the Foster-Seeley or phase discriminator. It consists of two Fig. 18.8*b* type a.m. detectors back to back, both receiving the incoming r.f. carrier-wave voltage *da* from a tuned circuit L_p C_p, and each receiving in addition half the voltage from the tuned circuit L_s C_s. L_p and L_s are the primary and secondary windings of an r.f. transformer, and because both windings are exactly tuned to the carrier wave and are coupled only very loosely the voltages across them are 90° out of phase, as indicated by the phasors da, ab and ac in Fig. 18.15*b*. The total r.f. input to diode D_1 is therefore *db*, and to diode D_2 it is *dc* (C_1 and C_2 being r.f. short-circuits). These inputs are equal, so the rectified voltages set up across the load resistors R_1 and R_2 will also be equal.

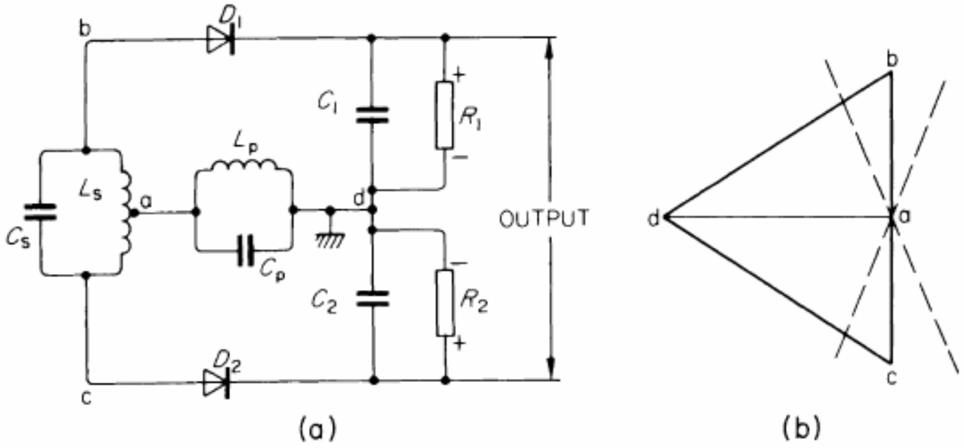


Fig. 18.15 (*a*) Circuit of phase discriminator for f.m. detection, arranged to bring out the principle: L_p and L_s are the primary and secondary windings of the input r.f. transformer; (*b*) phasors of voltages marked, for carrier wave frequency. Broken lines indicate the effect of f.m. deviation

But as they are connected in opposite polarity, the net output is nil.

When the carrier wave is frequency-modulated it swings off the resonance peak and, as explained in Sec. 8.11, the impedance of a tuned circuit changes from resistance to inductance on one side and capacitance on the other. The phase swings correspondingly, as indicted by the broken lines in Fig. 18.15*b*, changing the one-to-one ratio of *db* to *dc* to something alternately greater and less. The value of *db-dc* therefore becomes alternately positive and negative at the frequency of modulation, and the modulating signal is reproduced at the output. Practical phase-discriminator circuits differ in detail from Fig. 18.15 but the main principle is the same.

18.16 Ratio Detector

The ratio detector circuit combines the function of limiter with that of discriminator, so is usually preferred, in spite of slightly higher distortion. It looks very much like the phase discriminator, but one of the diodes is reversed, so when the carrier is unmodulated the total output from the diodes is not zero but is equal to the sum of the two. When modulation takes place this sum tends to remain constant but the ratio between the two rectifier outputs varies at audio frequency. The output of the detector can therefore be taken from either of the two reservoir capacitors as shown in Fig. 18.16. $R_1R_2C_1C_2C$ constitute the load circuit of a diode detector; thus the voltage across it (the sum of the voltages across C_1 and C_2) equals the peak r.f. input to the diodes, and the tuned circuit feeding the detector is effectively damped by a resistance equal to $(R_1 + R_2)/2$.

Ideally the r.f. input to the detector should be of constant amplitude but any noise signals present are superimposed on the f.m. carrier causing unwanted amplitude modulation. To reduce the

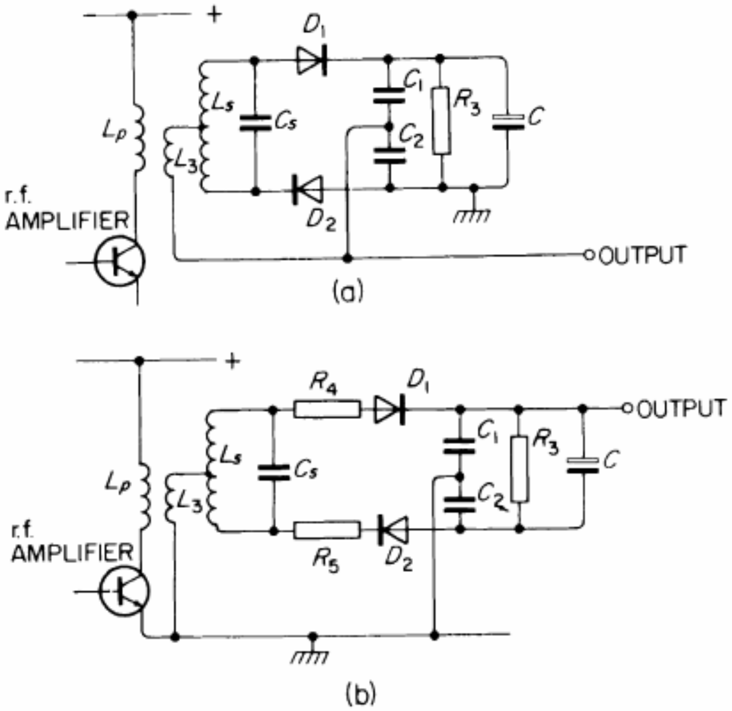


Fig. 18.17 Practical ratio detector circuits, (*a*) unbalanced and (*b*) balanced

input for the centre point of L_s is obtained from a coil L_3 tightly coupled to L_p. Circuit *a* is unbalanced, one side of the load circuit being earthed, and the output is taken from C_2. Circuit *b* is balanced, the midpoint of the diode load circuit being earthed. The output is taken from C_1 but it could equally well be taken from C_2. This circuit also includes resistors R_4 and R_5 to optimise the a.m. rejection. In both circuits the two load resistors R_1 and R_2 are replaced by a single equivalent resistor R_3. Notice, too, that C_p has been omitted; it is not essential to tune the primary circuit to the centre frequency.

18.17 Quadrature Detector

We now turn to a type of f.m. detector using entirely different principles. It is variously described as a balanced, symmetrical, product or quadrature detector, and its highly-complex circuit is encountered only in integrated circuit form; see Sec. 24.7. The principle is, however, straightforward and will be described with the aid of simplified diagrams. Two inputs are required with a particular phase relationship between them, and one method of obtaining suitable inputs is shown on the left in Fig. 18.18. One input is from the preceding r.f. amplifier and the other is derived from it by the network C_1LCC_2. C_1 and C_2 are very small capacitors and their reactance at the carrier frequency is large compared with the impedance of LC even at resonance. Thus the network is effectively capacitive and the current in it leads the applied voltage (input 1) by nearly 90°. At the resonance frequency of LC its impedance is signal. The energy content of the pulses depends on their duration, and the variation in energy can be turned into a variation in d.c. by passing the pulses through a low-pass filter to remove carrier-frequency components. The resulting signal is the required modulation-frequency output.

18.18 Phase-locked-loop Detector

This f.m. detector also uses different principles and again is a complex circuit encountered only in integrated-circuit form. As long ago as 1944 it was realised that if an oscillator could be synchronised by an f.m. signal so that its frequency faithfully followed the variations of the applied carrier, such a circuit could be expected to have two useful properties. Firstly the amplitude of the oscillator output could be many times that of the applied signal, implying a useful degree of voltage gain. Secondly the oscillator ouptut amplitude could be independent of that of the applied signal provided this is sufficient to give effective synchronisation: in other words the circuit should give effective amplitude-limiting. Thus the oscillator could be used as a source of amplified and limited f.m. signal which could then be applied to any type of f.m. detector. Used in this way, of course, the oscillator is not in itself a detector but a source of signal for a detector.

The phase-locked-loop circuit is a more recent application of the principle. It uses a phase discriminator, and a circuit of the type illustrated in Fig. 18.18 is generally used. But the second input is derived, not from the first input as in the quadrature detector, but from an oscillator forming part of the circuit. See Fig. 18.20. The oscillator is so designed that its frequency can be controlled by the value of a voltage applied to it. For example, the oscillator could be a Hartley or Colpitts type in which part of the capacitance of the LC circuit is provided by a capacitance diode, the d.v. input to which therefore controls the oscillation frequency. In the phase-locked-loop circuit the control voltage is the output of the phase discriminator and so we have a kind of negative feedback. Any variation in the frequency of the input signal causes an output from the phase discriminator which alters the oscillation frequency so as

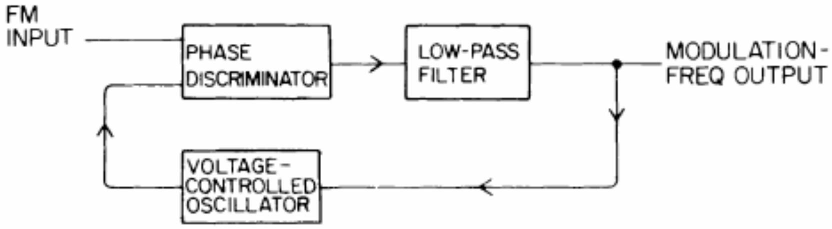


Fig. 18.20 Block diagram of a phase-locked-loop f.m. detector

to minimise the difference between its frequency and that of the input. In this way the phase of the oscillator is locked to that of the input and follows any variations in it. The output of the phase discriminator contains the required modulation-frequency component, but it is usually passed through a low-pass filter to minimise any r.f. components. This type of detector has the advantage that its performance does not depend on the amplitude/frequency characteristics of an LC circuit and can therefore be expected to be superior to that of detectors such as the Foster-Seeley, ratio and quadrature types.

18.19 Counter Discriminator

We have by no means exhausted the subject of f.m. detectors. For example there is the counter type, in which the cycles of the incoming f.m. signal are reduced to unidirectional pulses whose shape and duration are unaffected by changes in signal frequency. The number of them occurring per second, and therefore their average voltage, is, of course, directly proportional to the signal frequency. This linear relationship can be very accurately maintained, ensuring freedom from distortion, but the upper practical limit of carrier-wave frequency is only a few hundred kilohertz. As Chapter 21 will show, this is less of a handicap than might be supposed.

CHAPTER 19

Audio Frequency Amplification

19.1 Recapitulation

Amplification has already been considered in a general way and we have studied the aids used for designing amplifiers: characteristic curves and equivalent circuits. At the receiving end the r.f. signal picked up by the antenna usually needs amplification before being presented to the detector, which the last chapter has shown needs at least a volt or two if it is to do its job well. Besides amplifying, the receiver r.f. amplifier has to select the desired signals and exclude all others. And although it can be allowed to maltreat the waveform of the r.f. cycles, the waveform of the modulation must be carefully preserved. With f.m. this is no problem, but a.m. demands some care.

Our detailed study of transistor and valve parameters and equivalent circuits was for the most part based on neglecting reactances, so that phase relationships were just plus or minus. To what is called a first order of approximation, audio-frequency amplification can be treated on this basis, so we shall now proceed to do so, leaving receiver r.f. amplification (which certainly cannot be) until Chapter 22. Besides radio receivers, a.f. amplifiers are used in hearing aids, tape and disk recording and reproduction, and sound reinforcement.

The main problem is not in providing any desired amount of amplification; it is in doing so without perceptible distortion or extraneous noise. Perceptible, note. It is the human ear that judges the sound resulting from a.f. distortion, so it is important that the amount of distortion as well as the amount of amplification is reckoned as nearly as possible as the listener judges it. There are two main kinds of distortion: one we shall call *gain/frequency distortion** and occurs when the amplification or gain is not the same at all frequencies, so that sounds of different pitch are not reproduced at the correct relative strengths; the other is *non-linearity distortion* (or waveform distortion). In general it is more unpleasant than gain/frequency distortion. But first we should consider how amplitude and frequency are judged by hearers.

---

* B.S.I. calls this *attenuation distortion,* a most misleading term because it fails to indicate that the effect depends on frequency, and in this book we shall therefore use the self-explanatory term gain/frequency distortion.

19.2 Decibels

Suppose a 0.5 V signal is put into an amplifier which raises it to 12 V. This 12 V is then put into another amplifier which raises it to 180 V. Which amplifier amplifies more; the one that adds 11.5 V to the signal, or the one that adds 168 V? Put this way, the figures tend to suggest the second. But if we work out A, the voltage amplification ratio (Sec. 11.5) we find it is 24 for the first and only 15 for the second. This method of reckoning is the right one because we do not judge by the *amount* by which a sound is increased in strength, but by the *ratio* in which it is increased. Doubling the power into a loudspeaker produces much the same impression of amplification whether it be from 50 to 100 mW or from 500 to 1000.

This is like the way the ear responds to the pitch of sounds. While it is pleased by combinations of sounds whose frequencies have certain fixed ratios to one another, it can find no significance at all in any particular additive frequency intervals, say an increase of 250 Hz. Musical intervals are all ratios, and the significant ratios have special names, such as the octave, which is a doubling of frequency wherever on the frequency scale it may come. The octave is subdivided into tones and semitones.

The same principle is appropriate for the strength of sound or electrical signals, and amplification—or indeed any change in power—is reckoned in ratios. Instead of the 2 to 1 ratio, or octave, a power ratio of 10 to 1 was chosen as the unit. The advantage of this choice is that the number in such units is simply equal to the common logarithm of the power ratio. It was named the bel (abbreviation: B) after A. Graham Bell, the telephone inventor. So an amplifier that multiplies the power by 10 can be said to have a gain of 1 bel. A further power multiplication by 10, giving a total of 100, adds another bel, making 2 bels.

The bel, being rather a large ratio for practical purposes, is divided into 10 decibels (dB). One tenth of a 10 to 1 ratio is not at all the same as one tenth of 10; it is 1.259 to 1. If you have any doubts, try multiplying 1 by 1.259 ten times.

So a gain of 1 dB is a 25.9% increase in power. The voltage or current increase giving a 25.9% increase in power is 12.2% *provided that the resistance is the same for both voltages (or currents) in the ratio,* because power is proportional to the square of voltage or current (Sec. 2.17). The fact that a voltage or current decibel is a smaller ratio than a power decibel does not mean that there are two different sizes of decibels, like British and American gallons; it arises because the ratio of volt or amps is smaller than the ratio of power in the same situation. Decibels are fundamentally *power* ratios.

If 1 is multiplied by 1.259 three times, the result is almost 2, so a doubling of power is a gain of 3 dB. Doubling voltage or current quadruples the power, so is a gain of 6 dB. A table connecting decibels with power and voltage (or current) ratios is given in appendix E (p. 538). Since the number of bels is $\log_{10} P_2/P_1$, or 2 $\log_{10}$. V_2/V_1 (where P_2/P_1 is the power ratio and V_2/V_1 the voltage ratio) the number of decibels is 10 $\log_{10} P_2/P_1$ or 20 $\log_{10} V_2/\mathrm{V}_1$. If Fig. B.5*a* is a scale of dB, then *b* is the same scale in power ratios. The dB scale, being linear, is obviously much easier to interpolate (i.e., to subdivide between the marked divisions). Fig. 11.6 would be improved if the logarithmic voltage-ratio scale were replaced or supplemented by a uniformly divided scale with its 0 at 1 and its 20 at 10, which would indicate dB.

An advantage of working in dB is that instead of having to multiply successive changes together to give the net result, as with the *A* figures, they are simply added. A voltage amplification (A_V) of 24 is +27.6 dB, and $A_V = 15$ is +23.5 dB, so the total ($A_V = 360$) is +51.1 dB. The + indicates amplification or gain; attenuation or loss shows as a –. For instance, if the output voltage of some circuit were half the input, so that the power was one quarter—i.e., a power 'gain' of 0.25—the 1 to 4 power ratio would be denoted by the same number of dB as 4 to 1, but with the opposite sign: –6 dB. Zero dB, equivalent to $A = 1$, means 'no change'. The power losses in lines (Chap. 16) are reckoned in dB; if the line is uniform, the loss in dB is proportional to its length.

An important thing to realise is that decibels are ratios, not units of power, still less of voltage. So it is nonsense to say that the output from an amplifier is 30 dB, unless one has also specified a reference level such as 0 dB = 1 mW. Then 30 dB, being 1000 times the power, would mean 1 W.

To the general public, decibels are units of loudness, but they can be used as such only if a standard reference level is understood. If they only knew, dB can just as well be used for betting odds!

19.3 Gain/Frequency Distortion

The amount of gain/frequency distortion in any piece of equipment is shown by means of a graph in which gain (or loss) is plotted against frequency. For the reasons we have just been considering, both scales are usually logarithmic, the gain scale being linear in dB and the frequency scale usually running from 20 Hz to 20 kHz in three decades (i.e., 10:1 ranges).

Fig. 19.1 is a typical example. A curve having the same shape at a higher or lower level would mean the same amount of distortion. If a linear *A* scale were used the shape would vary with the level and so be misleading. An ideal a.f. gain/frequency curve would of course be flat or level over the whole range of frequencies to be reproduced.

As a guide to interpreting frequency characteristics, it is worth noting that 1 dB is about the smallest change in general level of sound

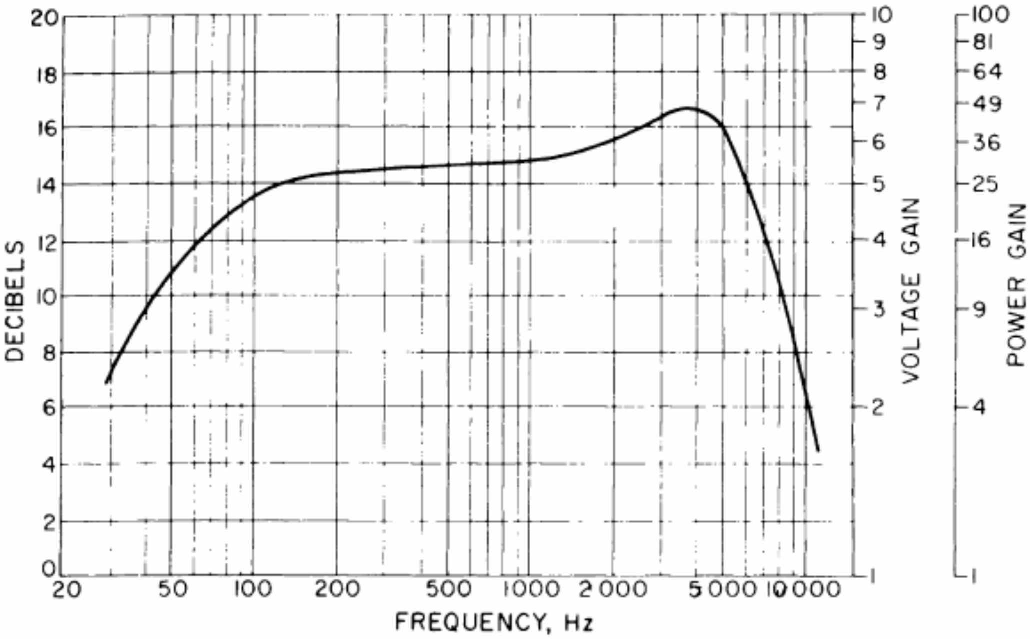


Fig. 19.1 A typical a.f. gain/frequency curve, with logarithmic scales in both directions

that is perceptible. If the loss or gain is confined to a small part of the audible frequency range, several dB may go unnoticed. So far as audible distortion is concerned, therefore, there is no sense in trying to 'iron out' irregularities of a fraction of a decibel. A peak of one or two decibels, though unnoticeable as gain/frequency distortion, may however cause non-linearity distortion by overloading the amplifier. An average fall of 10 dB over all frequencies above, say, 1 kHz would be easily noticeable as muffled and indistinct reproduction. A rise of the same amount, or a falling off below 1 kHz, would be heard as thin shrill sound.

A cheering thought is that gain/frequency distortion in one part of the system can usually be compensated in another by deliberately introducing distortion of the opposite kind, emphasising the weakened frequencies. But such methods, called *equalisation,* should not be relied on more than necessary, because they may cause difficulty in using one piece of equipment with others.

Means for the listener to adjust the frequency curve to suit his taste or to counteract distortion are called *tone controls*, and in modern hi-fi equipment can commonly be adjusted to give up to 12-15 dB gain or loss at the lowest and highest audio frequencies. They are usually labelled *bass lift, bass cut, top* (or treble) *lift* and *top cut.*

19.4 Non-Linearity Distortion

Non-linearity can be expressed as a graph of output against input. A diagonal straight line passing through the origin, as in Fig. 2.3, represents perfect linearity, and means that any given percentage increase or decrease of input changes the output in exactly the same proportion. Transistors are examples of non-linear devices, as shown by the curvature in their characteristics (Fig. 10.12, etc.).

Judging from an output-input graph, the most obvious effect of non-linearity might seem to be that the volume of sound produced would not be in exact proportion to the original. It is true that reducing the signal strength at the detector to one-tenth is likely to reduce the a.f. output to far less than one-tenth, because of the bottom bend (Sec. 18.11). But that is about the least objectionable of the effects of non-linearity. What matters much more is the distortion of waveform. Because the waveforms of a.f. signals correspond to those of the original sound waves, we must preserve their form with the utmost care. Otherwise the sound will be harsh and we shall not be able to distinguish one musical instrument from another (Sec. 1.3).

19.5 Generation of Harmonics

As far back as Fig. 1.2*c* we saw that when sine waves having frequencies in the ratios 1:2:3 etc. are added together the result is a non-sinusoidal waveform which repeats at its lowest frequency—the one represented by 1. In general the waveforms generated by musical instruments, buzzers, sirens, etc.—and, because of distortion, electronic oscillators—are more or less non-sinusoidal. It was shown by Fourier that repetitive waveforms can be analysed into sine waves having exactly multiple frequencies as above. These component waves are called *harmonics,* but the one of lowest frequency (which is also the frequency of the composite waveform) is usually called the *fundamental* rather than the first harmonic. The composite waveform depends on the number, amplitudes and relative phases of the harmonics. Let us now apply the idea of harmonic analysis to the output waveform of an amplifier.

Fig. 19.2 is a typical collector-current/base-voltage curve. If the load resistance were small compared with the output resistance of the transistor, the output voltage would be nearly proportional to I_C. Suppose P is the working point, from which a sinusoidal input signal swings the base between 0.4 V and 0.7 V. Plotting the output current waveform, we see that whereas the positive peak is 48 mA the negative peak is only 32 mA. This cannot, therefore, be a true sine wave, for which the peaks are equal. The average is 40 mA. In Fig. 19.3 a sine wave of that size has been plotted at *a*, a second harmonic having an amplitude equal to half the difference between average and actual peak output (4 mA) is shown at *b*, and *c* is the result of adding *a* and *b* together. It is almost identical with the output in Fig. 19.2. We conclude that characteristic curves of the Fig. 19.2 type, in

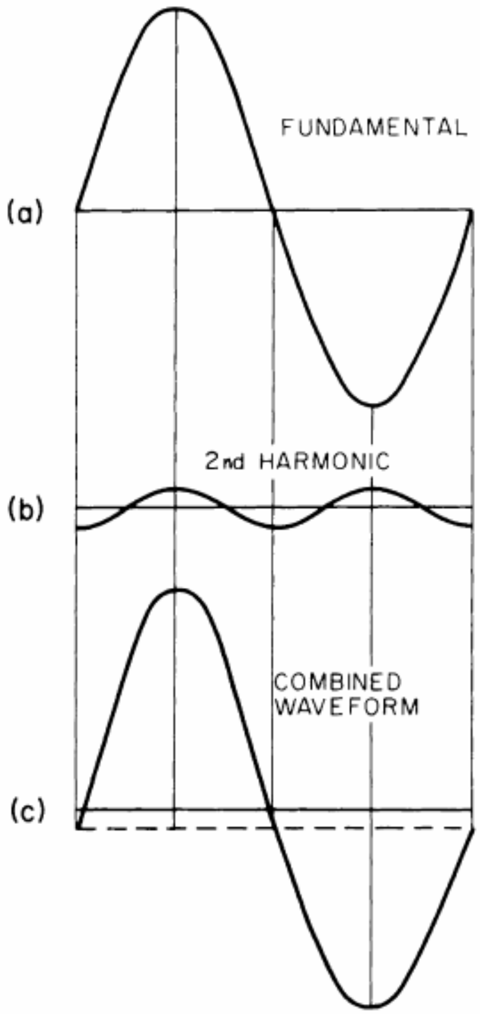


Fig. 19.3 Method of estimating the harmonic distortion in the Fig. 19.2 output

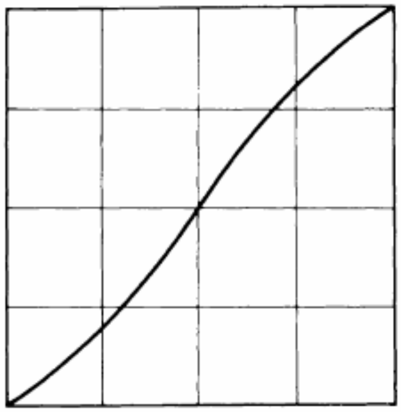


Fig. 19.4 In contrast to the type of characteristic curve in Fig. 19.2, which generates even-number harmonics, this double bend generates odd-number harmonics

of curvature, but usually one or other predominates. Unless there are sharp bends (which should always be avoided) the percentages of harmonics decrease as their order (i.e. number of times their frequency is greater than that of the fundamental) increases. Nevertheless, it is possible to select a particular harmonic by use of a tuned circuit resonant at that frequency; this is the basis of the *frequency multiplier.*

The ear tolerates a fairly large percentage of second-harmonic distortion, less of third harmonic, and so on. A fraction of 1% of the higher harmonics such as the eleventh introduces a noticeable harshness; the reason being that a second harmonic differs by an exact octave from the original tone, whereas the high harmonics form discords.

19.6 Intermodulation

Although for simplicity the distortion of a pure sine wave has been shown, the more complicated waves corresponding to musical instruments and voices are similarly distorted. So when one is dealing with waveforms other than sinusoidal it is usual to perform a harmonic analysis, and then tackle them separately on the basis of simple sine-wave theory; like the man in the fable who found it easier to break a bundle of faggots by undoing the string and attacking them one by one. Each harmonic current and voltage can be calculated as if it were the only one in the circuit—*provided that the circuit is linear.* But just now we are studying non-linear conditions, in which it is not true to say that two currents flowing at the same time have no effect on one another. We shall see in Sec. 21.3 that when two voltages of different frequency are applied together to a non-linear device the resulting current contains not only the original frequencies but also two other frequencies, equal to the sum and difference of the original two. For some purposes that is a very useful result, but extremely undesirable in an a.f. amplifier. It means that all the frequencies present in the original sound—and there may be a great many of them when a full orchestra is playing—interact or *intermodulate* to produce new frequencies; and, unlike the lower harmonics, these frequencies are generally discordant. Even such a harmonious sound as the common chord (C, E, G) is marred by the introduction of discordant tones approximating to D and F. When non-linearity distortion is severe, these intermodulation or combination tones make the reproduction blurred, tinny, harsh, rattling, and generally unpleasant. And once they have been introduced it is practically impossible to remove them.

19.7 Allowable Limits of Non-Linearity

Whenever one attempts to get the greatest possible power from a transistor or a valve, one comes up against non-linearity. So the problem tends to be most acute in the output stage. If the signal amplitude is reduced, so that it operates over the most nearly linear parts of the characteristics, the distortion can be made very small, but then the output will be uneconomically small in relation to the power being put in. It is not a question of adjusting matters until distortion is altogether banished, because that could be done only by reducing the output to nil. It is necessary to decide how much distortion is tolerable, and then adjust for maximum output without exceeding that limit.

Fixing such a limit is no simple matter, because it should (in reproduction of music, etc.) be based on the resulting unpleasantness judged by the listener. Different listeners have different ideas about this, and their ideas vary according to the nature of the programme. And, as has just been pointed out, the amount of distortion that can be heard depends on the order of the harmonic (second, third, etc.) or of the intermodulation tones, as well as on the percentage present.

Various schemes for specifying the degree of distortion in terms that can be measured have been proposed, but the nearer they approach to a fair judgment the more complicated they are; so the admittedly unsatisfactory basis of percentage harmonic content is still commonly used. Preferably the percentage of each harmonic is specified, but more often they are all lumped together as 'total harmonic distortion'.

19.8 Phase Distortion

The main result of passing a signal through any sort of filter, even if it is only a resistance and a reactance in series, is to alter the amplitudes of different frequencies to different extents (Fig. 18.9). If it discriminates appreciably between frequencies within the a.f. band it introduces gain/frequency distortion, which, of course, may be intentional or otherwise. It also shifts the phases to a differing extent; see Fig. 6.10 for example. If the signal is a pure sine wave, the phase shift has no effect on the waveform; but few people spend their time listening to continuous pure sine waves.

To take a very slightly more complex example, a pure wave is combined with a third harmonic, shown separately in Fig. 19.5*a*. The waveform of this combination, obtained by adding these two together, tends towards squareness, as shown at *b*. Suppose now that after passing through various circuits the relative phases of the component waves have been altered, as at *c*. Adding these together, we see the resulting waveform, *d*, which is quite different from *b*. Yet oddly enough the ear, which is so sensitive to some changes of waveform (such as those caused by the introduction of new frequencies), seems quite incapable of detecting the difference between *d* and *b*, or indeed any differences due to phase shifts in continuous waves.

Generally, it can be said that if care is taken to avoid phase distortion, gain/frequency distortion will take care of itself.

19.9 Loudspeakers

Having considered distortion we begin now to apply this knowledge to a.f. amplifiers. Their end product is usually sound from a loudspeaker, which forms the load for the final or output stage of the amplifier. Until we know the requirements here we cannot tell how many and what stages will be needed between the a.f. signal source—detector, tape head, microphone, etc.—and the output. So let us begin at the end with the loudspeaker. Its object, of course, is to convert the electrical power delivered to it from the output stage into acoustical power having as nearly as possible the same waveforms.

them, together with a constant p.d. of the order of a thousand volts which creates a steady attraction. The signals generate variations in this, causing the sound reproduction, which can be the best of any due to absence of resonances, and the wide range of audio frequencies, especially at the high end, reproducing transients especially faithfully. The thin, flat shape is pleasing too. So it is favoured by many hi-fi enthusiasts. But its undoubted merits are at the expense of high price, the provision of the high polarising d.v., low efficiency, and awkward load characteristics.

19.10 The Output Stage

In considering how much power can be got out of a transistor or valve, regard has to be paid to the maker's maximum ratings. These

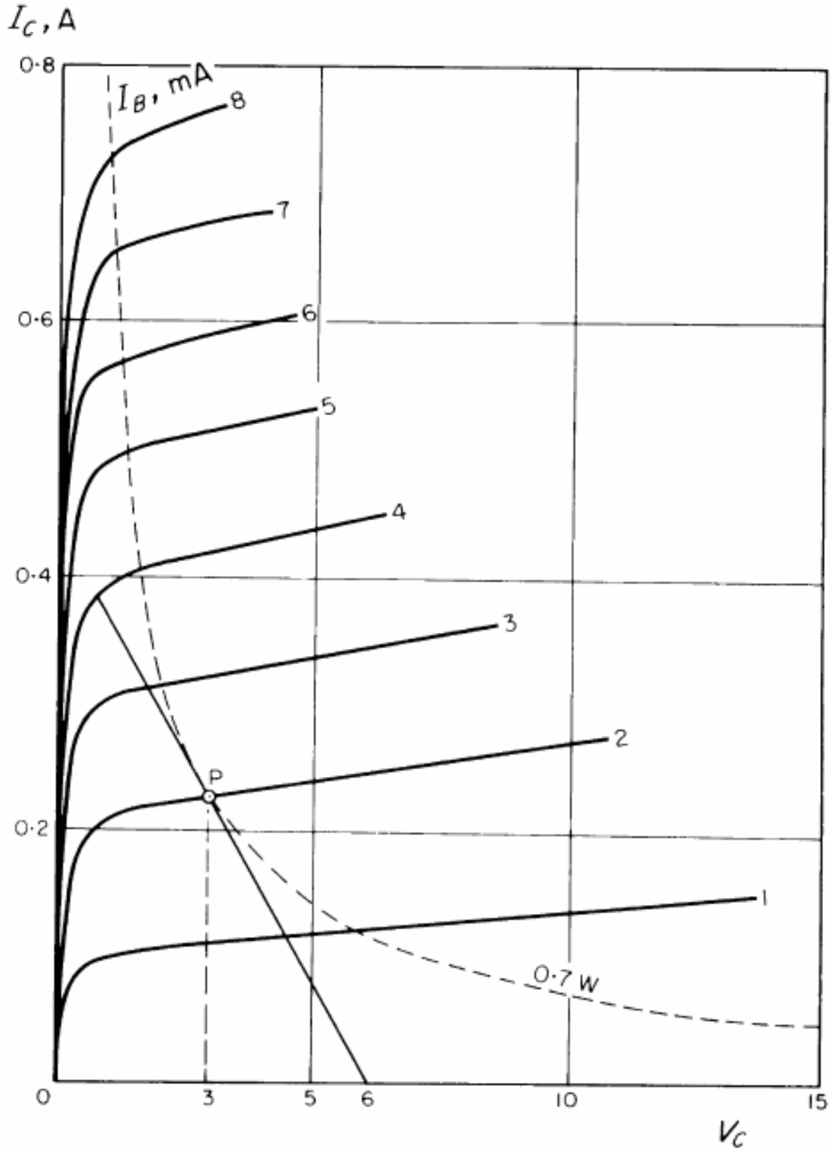


Fig. 19.7 I_C/V_C curves of a small a.f. output transistor, showing power-limit curve (dashed) and Class A load line

load resistance should be made double that suitable for one.

While the heavy power consumption of Class A amplifiers may perhaps be tolerable in amplifiers worked from the mains, and at least has the merit of being constant, it is quite intolerable when batteries are used, their cost per unit being enormously greater. The particularly undesirable feature is the continuance of full power consumption during quiet periods, so that the average efficiency is very low indeed.

19.11 Class B Amplification

The inadequate power output indicated when the transistor curves in Fig. 19.7 are examined on a Class A basis prompts us to look into other methods of operation. Class C however is out, because the half-cycles are so distorted. In Class B the amplifying transistors are biased close to cut-off, so that the quiescent power consumption is nearly zero, yet yield complete half-cycles for putting together to make whole ones. At all volumes—not only at maximum—most of the power expended is returned usefully. Translating this ideal into practice has its problems of course, as we shall see.

The load-line procedure we used with Class A does not work with Class B, because the working point is near cut-off, where the power input is very small. In fact, it is usually referred to as the quiescent point. And even when a load line has been drawn it does not directly indicate the power which the transistor will be called upon to dissipate. So we must try a different approach.

Fig. 19.8 shows the basic circuit diagram and ideal characteristics with zero saturation voltage and no bottom bend. So the quiescent

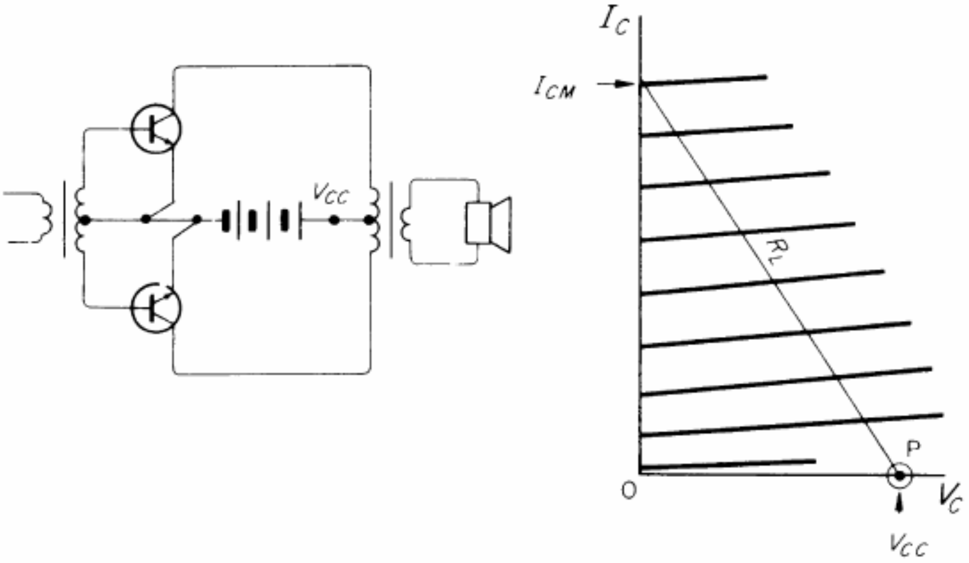


Fig. 19.8 Basic circuit diagram and ideal transistor characteristics for Class B output stage

cycles, instead of one being cut off, so the whole-primary resistance is effectively $2R_L$.

In Fig. 19.9 a more practical load line is shown, though still with zero quiescent current, on the same curve sheet as Fig. 19.7. The resistance it represents is 15Ω, and the signal peak is 11 V 0.72 A, giving an output power per transistor of practically 2 W. The input is $12 \times 0.72/\pi$, = 2.75 W, so with this type of signal the dissipation is just over the rated 0.7 W which compares with the calculated value of 0.5875 W for ideal conditions.

19.12 Distortion with Class B

In small-signal transistors the I_C/I_B relationship (h_{fe}) is remarkably linear, as we saw in Fig. 10.11, whereas I_C/V_B is very non-linear because the working range is near the bottom bend. Power transistors on the other hand are worked high up the input curve, where I_C/V_B tends to be linear and I_C/I_B shows a marked falling off as I_C increases. This is noticeable in Fig. 19.9, where the spaces between the I_B lines get progressively smaller as one moves up the

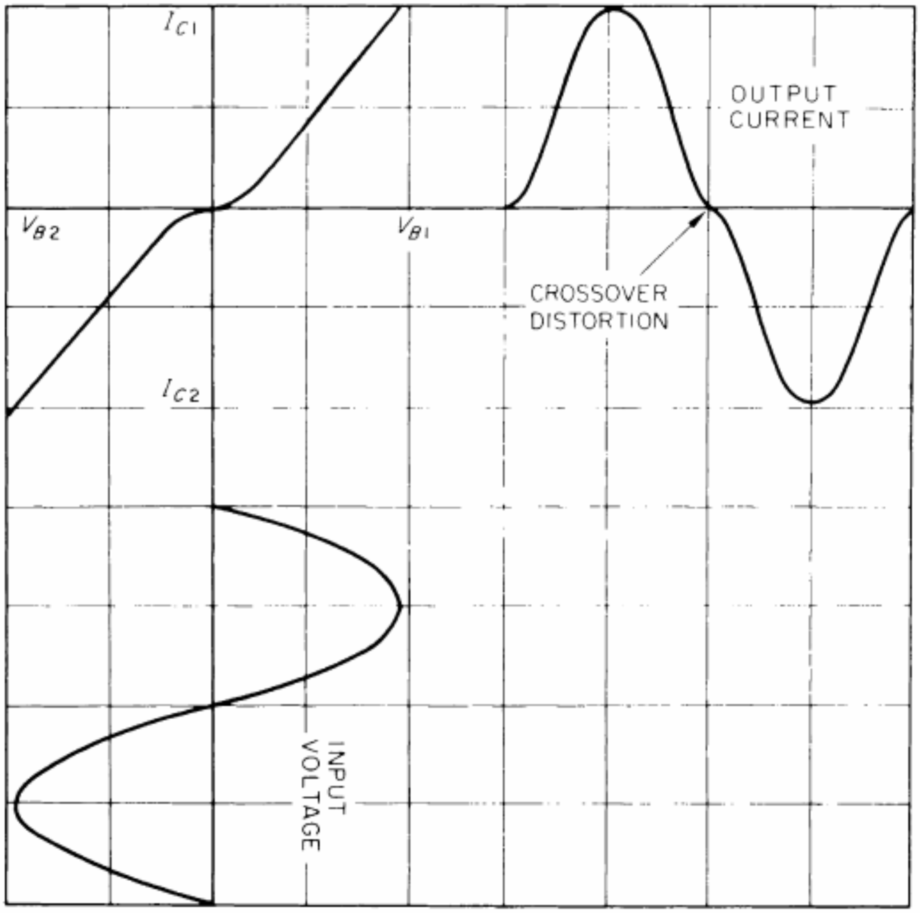


Fig. 19.10 Showing the production of crossover distortion in a Class-B output stage

The remedy is to bias the transistors to about the middle of the bottom bends, as in Fig. 19.11. If the biasing has been well judged the two output waveforms will add up to a very acceptable enlarged copy of the input. The crossover region is shown inset on a larger scale. Obviously the success of this method depends on how accurately the biasing has been set in the first place and kept there in spite of changes in transistor characteristics, due mainly to temperature changes. It is not much good having a beautifully undistorted output immediately after switching on if severe distortion sets in when the transistors have warmed up. This is where we must look at actual circuits.

19.13 Class B Circuits

What might be called the conventional push-pull circuit is as shown in Fig. 19.8; whether it is Class A, B or C (or one of their hybrids) depends on the amount of bias used. Unlike the tuned r.f. input and output transformers used in r.f. amplifiers, a.f. transformers are designed to give uniform results over the whole a.f. band (or as much of it as the designer can manage for the money). The input transformer is needed to apply positive signal to one transistor while the other is receiving negative signal. In Class A both are working all the time, each delivering full cycles; in Class B each works half time, alternately delivering half cycles. An output transformer is needed to combine the outputs in the correct phase to give to the load. With some transistors it is also needed to match their high output resistance to the low loudspeaker impedance. As we have seen, the most suitable load resistances for transistors are often in the same range as those of loudspeakers. A.f. transformers—and especially output transformers—are heavy and expensive if they are to be good, and if they are not good they cause both distortion and loss, so it seems a pity to use one merely for combining the outputs from the two transistors. And it is in fact unnecessary to do so.

If Class B transistors are fed in series, as in Fig. 19.12, instead of in parallel as in Fig. 19.8, the two ouputs flow through the load

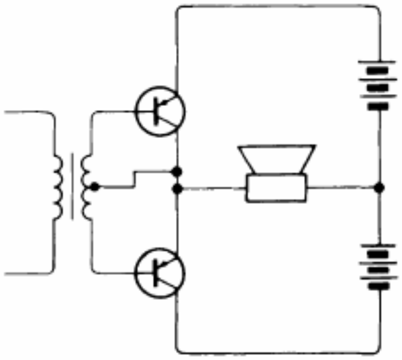


Fig. 19.12 One form of Class B output circuit. Its disadvantage is the centre-tap on the power supply

In this particular formation, with the input returned to either b or c, both transistors are working in common-collector mode, so the input signal voltage would have to be actually greater than the output. To avoid this disadvantage by using the common-emitter mode it would seem that the input should be returned to the emitters, point a. But as this point is alternating at the full output signal voltage the driver would have to be coupled to the output stage by a transformer, which is what we are trying to avoid. A way out of this difficulty is shown in Fig. 19.15, where the input voltage is developed across R_1, the top end of which is returned to output emitters signalwise via C, while the required d.c. feed is obtained via the loudspeaker. The two output transistors receive the usual current-adjusting emitter bias by means of R_2 and R_3, which because of the large currents flowing through them are typically only an ohm or two. Base bias is supplied by the voltage drop across D (Sec. 13.7). This voltage changes comparatively little over quite a wide range of current. One silicon diode usually provides enough voltage (about 0.6 V) for a pair of germanium transistors; if they are silicon, two diodes in series will be needed.

19.14 Compound Transistors

A convenient method of obtaining a higher current gain than is provided by a single transistor is by combining two as shown in Fig. 19.16, forming what is knownas a compound pair, or Darlington pair. The essential feature is that the emitter current of the first transistor is the base current of the second, so the current gain of the pair is equal to the product of the current amplification factors of the two transistors; i.e., $h_{fe1} \times h_{fe2}$. This result is achieved with a minimum of components, so the idea is used for many purposes besides power amplifiers.

Biasing arrangements are similar to those for one transistor, but of course the actual bias voltage required—the difference between V_{B1} and V_{E2}—is equal to the sum of the two separate bias voltages.

In Fig. 19.16 the pair is shown in common-emitter mode. When used in common-collector mode the feedback effect (Sec. 12.9) also is multiplied, so that the input resistance is exceptionally high.

19.15 Negative Feedback

The effect on the apparent resistance of a circuit of injecting into it a voltage or current proportional to the voltage or current applied was explored in Chapter 12, where in most cases the injected signal was in opposition to the applied signal. In amplifiers this is called negative feedback. It is often used intentionally to increase or reduce the input or output resistance of an amplifier. It is also much used for comparison the input scale has been reduced to one tenth, as if a distortionless 20 dB amplifier had been added before the input. Also for clarity the curve has been shifted to the right, and the corresponding input swing around P′ as working point is from A′ to B′. The positive and negative output half-cycles are now much more nearly equal than in *a*. The second-harmonic distortion is too small to read off the curve, but if calculated would be found to be one tenth as much as in *a*.

Suppose now that the input is increased by 50% so as to include the dotted portions. In contrast to the practically linear D′A′, the bit from A′ to C′ would cause a more drastic increase in distortion than AC because it contains a more acute change of slope where it meets the axis. So one has to be careful, when using an amplifier with negative feedback, not to push the volume control too far.

What about gain/frequency distortion? Suppose that in the example just considered the gain falls off at the highest a.f. to the extent of 6 dB, or 50%. At the lower frequencies, where the full gain operates, the reducing factor of the feedback is $1/(1+9)=0.1$ as we assumed. When A is halved this is $0.5/(1+4.5)=0.091$. So the relative gain at the top frequencies, with feedback, is $0.091/0.1$, $=0.91$, or -0.8 dB, which is hardly noticeable. Negative feedback therefore tends to flatten out irregularities in the gain/frequency characteristic.

19.16 Phase Shift with Feedback

The main cause of gain/frequency distortion in amplifiers is reactance. In Sec. 18.10 we considered the use of a resistance-capacitance potential divider as an elementary sort of filter for removing r.f. signals from the output of a detector. Either kind of reactance causes an increasing loss at either high or lower frequencies, depending on where it comes in relation to resistance; both kinds of reactance together can cause either a loss or gain around one particular frequency, since they form a tuned circuit. In Sec. 6.7 and elsewhere this frequency discriminating effect was shown to be accompanied by a phase shift. So gain/frequency distortion is accompanied by phase distortion. As negative feedback reduces gain/frequency distortion it seems reasonable to infer that it reduces phase distortion too. And so it does, but with important reservations as we shall see.

For manufacturing reasons, inductance tends to be avoided in modern a.f. circuits. Where a.f. transformers are still used they are analysed into their equivalent circuits (Sec. 7.10) for calculating frequency effects, which can be done by drawing phasor diagrams for a number of typical frequencies. There is inevitably some inductance in loudspeakers, some of it purely electrical and some

19.17 Input and Output Resistance

The theory of the effects of feedback on input and output resistance is given in Sec. 12.5; now we can glance at it again with particular reference to amplifiers. Usually we would be glad to increase the input resistance of a bipolar transistor so as to make it as nearly as possible voltage-operated, more like an f.e.t. or valve. We saw in Sec. 12.9 how a great increase is obtainable by working it in common-collector (or emitter-follower) mode, and in Sec. 19.14 a further increase by using a pair of transistors in this way. The effect is produced by what amounts to total negative feedback; *all* the output voltage is fed back to the input. So the input voltage to cause a given input current must be increased $(1 + A)$ times and the apparent input resistance is therefore $(1 + A)$ times greater.

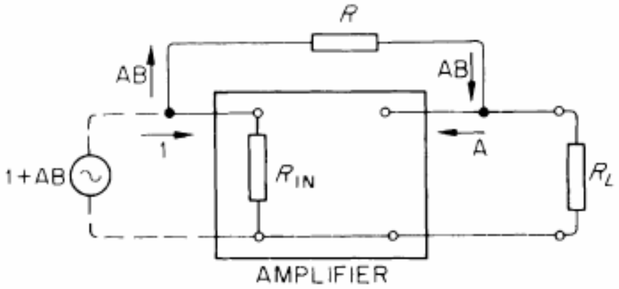


Fig. 19.22 Showing the basic principle of shunt-injected negative feedback

In Fig. 19.17 some of the output voltage was fed back in series with the input. An alternative method in which output signal is fed back in parallel is shown in Fig. 19.22. It is sometimes called shunt-injected feedback. Treating the system as a current amplifier we again call the input 1, this time say 1 μA. With a current amplification A, there is an output of A microamps in the direction shown. Of this, a fraction B, controlled by R, is fed back to the input. (The fact that the arrows seem to show it being fed forward is because it is *negative* feedback.) So for every microamp going into the amplifier the signal current source has to supply $(1+AB)$ microamps. The current gain (neglecting the fraction of output bypassed via R) is therefore reduced to $A/(1+AB)$ as in Fig. 19.17, and because the current taken from the source is $(1+AB)$ times as much as is actually going into the amplifier, the input conductance seems to be $(1+AB)$ times as great, and the resistance is that much less.

Output resistance, as we saw in Sec. 12.7, is measured by substituting a signal generator for the load and short-circuiting the input voltage generator, the only input then being what is fed back. This is not difficult to do theoretically by calculation, and the result is that voltage feedback reduces the apparent output resistance of an amplifier. But the extent of this effect is greater than the factor injected into the input circuit. If it is connected in parallel with the input, as in Fig. 19.22, it reduces the effective input resistance but if it is injected in series with the input (as by omitting the emitter bypass capacitor) it increases the input resistance.

When a short-lived signal (called a *transient*) having a frequency at which the loudspeaker it is fed into resonates, the loudspeaker continues to vibrate and emit sound after the signal has ceased. This effect is called *transient distortion*. The oscillation of the moving coil in the strong magnetic field (Sec. 19.9) generates an e.m.f., and as it is in parallel with the output of the amplifier it will make a current flow therein. How much current depends on the output impedance. If it is very small—say a fraction of the loudspeaker's own impedance—the current will be correspondingly large and will quickly damp out the undesirable oscillation. This is obviously a good thing, which can be achieved by quite a moderate amount of negative voltage feedback.

The 100% negative feedback in emitter followers and cathode followers not only greatly increases the input resistance but reduces the output resistance. So although the voltage gain is less than 1 the current gain can be very large. These devices are therefore largely used as a sort of current transformer to feed low-impedance loads from high-impedance sources. They have two great advantages over inductive transformers: the voltage loss is very small; and the effective frequency band can be very large, including zero frequency. Fig. 19.22 can be adapted to illustrate 100% negative feedback by transferring R_L to the R position so that all the output signal is fed back.

19.18 Linearising the Input

The advantages of negative feedback in a.f. amplifiers are so valuable that it is practically always used. The number of ways in which it can be applied are too many to show here, but armed with the foregoing principles we ought to be able to follow most of the circuit diagrams we see.

Many of the low-power transistors used in the initial stages of amplifiers have very linear I_C/I_B characteristics, as in Fig. 10.11, and very non-linear I_C(or I_B)/V_B, as in Figs. 10.12 and 13. Obviously they should be worked from signal current sources rather than voltage. Where the impedance of the actual source was not sufficiently high to qualify it as a current source, a common practice was to bring it up to that standard by adding resistance. Fig. 19.24 shows an input-stage circuit (feeds omitted) in which the transistor is assumed to have a perfectly linear I_C/I_B curve (i.e., constant h_{fe}). With the voltage signal source S connected direct there would be serious distortion because of the very non-linear input resistance

(I_B/V_B curve). Inserting a relatively large linear R makes the total resistance nearly linear and therefore the input current, and therefore also the output current, nearly proportional to the input signal voltage. But of course most of the signal is lost in R.

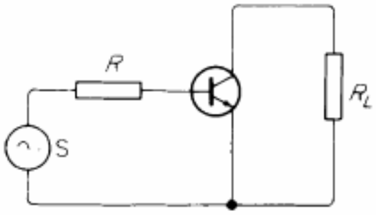


Fig. 19.24 One way of linearising a transistor is by means of a relatively high resistance R in series with the input

If instead negative voltage feedback were used the result would be just the same; the voltage drop due to I_B flowing through R would be replaced by a voltage drop due to $h_{fe}I_B$ flowing through part of R_L. But whereas the series-resistance method linearises only the input circuit, and depends on linear I_C/I_B (which is now less often found), feedback can be applied to as much of the amplifier as one wants and corrects all non-linearities, so has virtually superseded series resistance.

19.19 Some Circuit Details

To gather together the fragmentary studies of this chapter and to illustrate some further features, a complete a.f. amplifier circuit is shown in Fig. 19.25. Typical component values are given only to

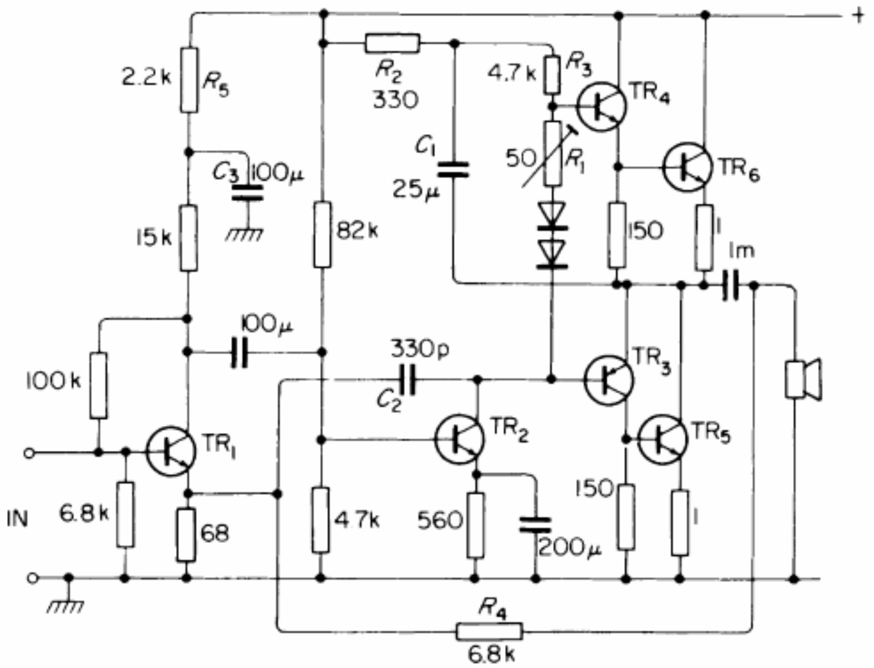

+
2.2k
R_5
R_2
330
4.7k
R_3
TR_4
C_1
25μ
50
R_1
TR_6
C_3
100μ
15k
82k
150
1
1m
100μ
330p
TR_3
100k
C_2
TR_2
TR_1
TR_5
IN
6.8k
68
4.7k
560
150
1
200μ
R_4
6.8k

Fig. 19.25 A typical complete audio amplifier circuit

input is progressively increased, the class-B stage comes increasingly into operation and, at full output, supplies most of the power to the loudspeaker. Because of the absence of standing forward bias, the class-B stage introduces considerable crossover distortion which can, however, be virtually eliminated by the use of the two negative feedback loops shown in dashed lines, one embracing the class-A stage and the other the entire amplifier. The class-B stage is regarded as a current dumper*; alternatively the whole amplifier can be regarded as an example of a process known as 'feed-forward error compensation,' the class-A stage supplying the output at low signal levels where the dumper stage cannot operate satisfactorily; i.e., the class-A stage corrects the dumper-stage error.

19.20 Noise

While the gain of an amplifier such as the one in Fig. 19.25 is sufficient to give full output from a radio detector or some types of record player pick-up, it is usually not enough for the better types of pickups and microphones or for tape playback heads. Amplifiers for many other purposes also have to deal with even smaller signals, perhaps only a few microvolts. Designing for the required gain is easy; the real problem is *noise.* This word is used to refer to electrical disturbances which, in sound reproducers, cause audible noise; the same disturbances are also objectionable when the end product is not sound—they mar the picture in television, for example.

Some noise is picked up from external electrical equipment. There is hum from the public a.c. supply. And anything that causes abrupt changes of current generates noise over a wide frequency band; such things as switching, thermostat contacts, appliances using commutator motors, and faulty lighting. Such disturbances can be excluded from a.f. amplifiers by properly arranged earthing, electric and magnetic screening, filters and careful layout of wiring. Printed wiring and integrated circuits (Sec. 24.7) are less likely to pick up induced noise.

But when all such noise has been suppressed or excluded there is a never entirely avoidable residue generated in the amplifier circuits themselves. An electric current is not a smoothly continuous flow; it is made up of individual electrons, and even when there is no net current in any one direction the electrons in, say, a resistor are in a continuous state of agitation. Their random movements, like the aimless jostling of a crowd which as a whole remains stationary, are equivalent to tiny random currents, which give rise to small voltages across the resistor. Quite a moderate gain, such as can be obtained there is what is called *flicker noise,* or 1/*f* noise because it is inversely proportional to frequency.

---

*An unfortunate term suggesting that the current of this stage is thrown away uselessly, whereas at high signal levels it supplies most of the wanted output.

The noisiness of a device or stage, or even of a whole amplifier or receiver, is specified as its *noise factor* or noise figure, which is defined in various ways, of which a simple one is

$$\frac{\textit{input signal/noise power ratio}}{\textit{output signal/noise power ratio}}$$

It is usually expressed in dB; 0 dB is perfection, and the higher the figure the worse.

19.21 Another Box Trick

In Sec. 12.5 we introduced a box with two terminals whose electrical properties were not what would be expected from its contents. Here is another example of such a box. This time a sinusoidal voltage applied to the terminals gives rise to an input current in quadrature with the voltage. We can therefore say that the box behaves as a reactance—capacitive if the current leads the voltage, and inductive if it lags. But the box contains no inductance or capacitance to provide the reactance, the box behaviour being brought about by a common-emitter amplifier with a phase-shifting negative feedback loop.

A typical circuit of this type is shown in Fig. 19.27. Suppose C is given a value such that its reactance at the operating frequency is small compared with R. The current in RC is then in phase with the collector voltage. The voltage across C lags this current by 90° and so also does the base signal and the collector current. Thus the collector current lags the collector voltage by 90° and the collector/emitter path of the transistor behaves as an inductor. The magnitude of the inductance can be calculated in the following way.

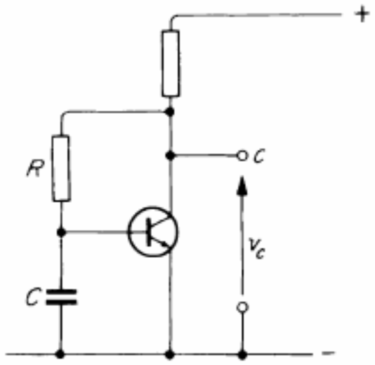


Fig. 19.27 A simple reactance stage in which the transistor behaves as an inductor

If the collector voltage is v_C, the current in RC is given approximately by v_C/R and the voltage across C is $v_C/R\omega C$. The collector current is thus $g_m v_C/R\omega C$ where g_m is the mutual conductance of the transistor. If we represent the effective

CHAPTER 20

Selectivity and Tuning

20.1 Need for R.F. Amplification

In Sec. 18.11 we saw how the effectiveness of detectors deteriorates quite rapidly as the r.f. input is reduced much below about 1 V. Signals picked up by a receiving antenna are seldom more than a few millivolts and are often only microvolts, so (except from a powerful sender at short range) they are too weak to operate a detector satisfactorily. Amplification of the r.f. signals is therefore a practical necessity in order to give a receiver adequate *sensitivity,* which, for an a.m. sound receiver, is usually expressed as the input microvoltage needed to yield a specified a.f. output (often 50 mW) at a specified modulation depth (usually 30%).

Another practical necessity is *selectivity*, which as we saw in Sec. 8.6 is the ability of a receiver to discriminate in favour of the wanted signal against all others, some of which may be much stronger and only slightly different in frequency. As will have been gathered, the chief means for obtaining selectivity is the tuned circuit. Although tuned circuits are not necessarily associated with amplification they almost invariably are, for reasons that will appear. Meanwhile we shall do well to make a closer study of selectivity.

20.2 Selectivity and Q

We saw in Chapter 8 that both selectivity and sensitivity are improved by increasing the Q of a tuned circuit. Fig. 8.5, for example, shows how the response at the frequency to which a circuit is tuned increases in direct proportion to Q, whereas the response at other frequencies is less affected; in fact, at frequencies far off tune it hardly changes at all. The problem of r.f. amplification might therefore seem to be simply one of getting a high enough Q. We found (Sec. 8.4) that the Q of a circuit is the ratio of its reactance (either inductive or capacitive) to its series resistance, which is the same thing as the ratio of its dynamic resistance to its reactance:

$$Q = \frac{X}{r} = \frac{R}{X}$$

Now the amount of reactance is decided by the requirements of tuning, so attention must be confined to reducing the series resistance, r. This will have the effect of increasing the dynamic resistance, R. We noted (Sec. 8.14) that r is a sort of hold-all figure

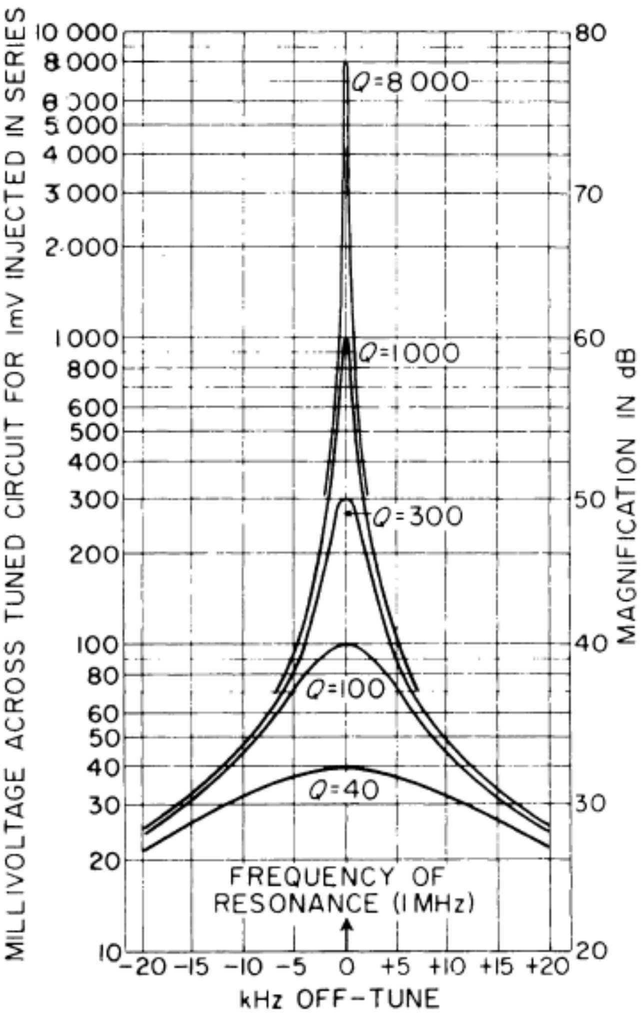


Fig. 20.1 Showing how r.f. voltage across a tuned circuit varies with frequency at different *Q* values, obtained by adjusting positive feedback

desired information (using that word in its widest sense). If the carrier is frequency-modulated (Sec. 15.4) it is obvious that a definite band of frequencies, with the carrier frequency at its centre, must be received to an equal degree. The need for covering a band of frequencies when the carrier is amplitude-modulated is less obvious, but it none the less exists.

20.3 Over-Sharp Tuning

If we look back at Fig. 20.1 we see that the curve for $Q = 8000$ fails conspicuously to meet this requirement even for a.f. modulation (Sec. 15.8) and does not begin to meet the much more demanding needs of video modulation (Sec. 15.9). The sidebands corresponding to a modulation frequency of 5 kHz are shown to be received only to the extent of 1.3% as much as the carrier wave and the lowest a.f. sidebands. The high notes are quite literally tuned out by over-selectivity. Higher pitched sounds still would be even more drastically attenuated. The result would be indistinct speech and boomy music. Even with a Q as low as 100, 30%, or 3 dB, of 5 kHz modulation would be lost, but this would not be too much to restore by an equaliser (Sec. 19.3).

Because the effective frequency bandwidth for a given Q is proportional to the carrier frequency, whereas the bandwidth required is unaffected by it, the sideband cutting effect is less at higher carrier frequencies and greater at lower carrier frequencies.

With frequency modulation there is the same need to preserve sidebands, but the results of failing to do so are different. Fig. 15.7 might certainly suggest that if the bandwidth were limited to 150 kHz all except perhaps the lowest audio frequencies would be lost. They would however all be represented at least to some extent by the sidebands up to 150 kHz. But this graph represents 100% modulation, whereas most of the time the percentage is much less. The main effect of sideband cutting is therefore to limit the depth of modulation—and hence the resulting a.f. amplitude—that can be correctly reproduced. In other words, non-linearity distortion, in contrast to gain/frequency distortion of a.m. signals.

This undesirable effect is at least partly counteracted by the limiter in an f.m. receiver, which tends to bring all r.f. signal amplitudes down to one level, but it could not cope with really severe sideband cutting.

Any hopes we may have entertained of achieving adequate sensitivity and selectivity quite simply by pushing up Q are therefore bound to fade.

20.4 A General Resonance Curve

Whether we are considering channel separation (Sec. 15.9) or loss of sidebands, then, the basis of reckoning is 'frequency off-tune'. To distinguish this particular frequency from others we shall denote it by f'.

The ideal shape for a tuning curve would seem to be one which was level over a frequency band wide enough to include all the wanted sidebands and then cut off completely to exclude everything else. None of the resonance curves we have seen so far looks anything like this—see Fig. 20.1, for example. And except for Fig. 8.6 they are arranged for comparing sensitivity rather than selectivity. We need to take the response at resonance as the starting point and see how much it falls at off-tune frequencies. This falling-off can be reckoned in dB. An alternative scale for any who may still have difficulty in thinking in dB is in what we shall call S numbers, which are the numbers of times by which response off-tune is less than that at resonance. We now want to know what determines S at any given

20.5 More Than One Tuned Circuit

Let us now consider what happens when additional tuned circuits are used. If we can assume that these circuits do not react on one another, then it is simple. If one circuit reduces signals at a certain off-tune frequency 3 times (relative to the response at resonance), and another circuit is somewhat more lightly damped and reduces them 4 times, then the total relative reduction is 12 times. If a third circuit gives a further reduction of 3 times, the total is 36. In other words, an r.f. amplifier incorporating these three tuned circuits amplifies this certain frequency 36 times less than it amplifies the frequency to which all three circuits are tuned. In decibels, the three circuits cause reductions of 9.5, 12 and 9.5 dB; the total is found by simple addition: 31 dB.

If the problem is simplified by assuming all the circuits have the same L/r, then the curve in Fig. 20.3 is made to apply to a pair of circuits by squaring S; to three circuits by cubing it; and so on. Or by multiplying the dB scale by 2, 3, etc.

Comparison between one and several tuned circuits can be made in different ways. First of all we shall compare them on the basis of equal gain, in order to dispose finally of the question of positive feedback versus r.f. amplification. Taking some typical figures, suppose that the r.f. amplifier is a single stage giving a gain of 50 times, with input and output tuned circuits for which $L/r = 10$. Squaring the S values in Fig. 20.3, we get curve *a* in Fig. 20.4. To

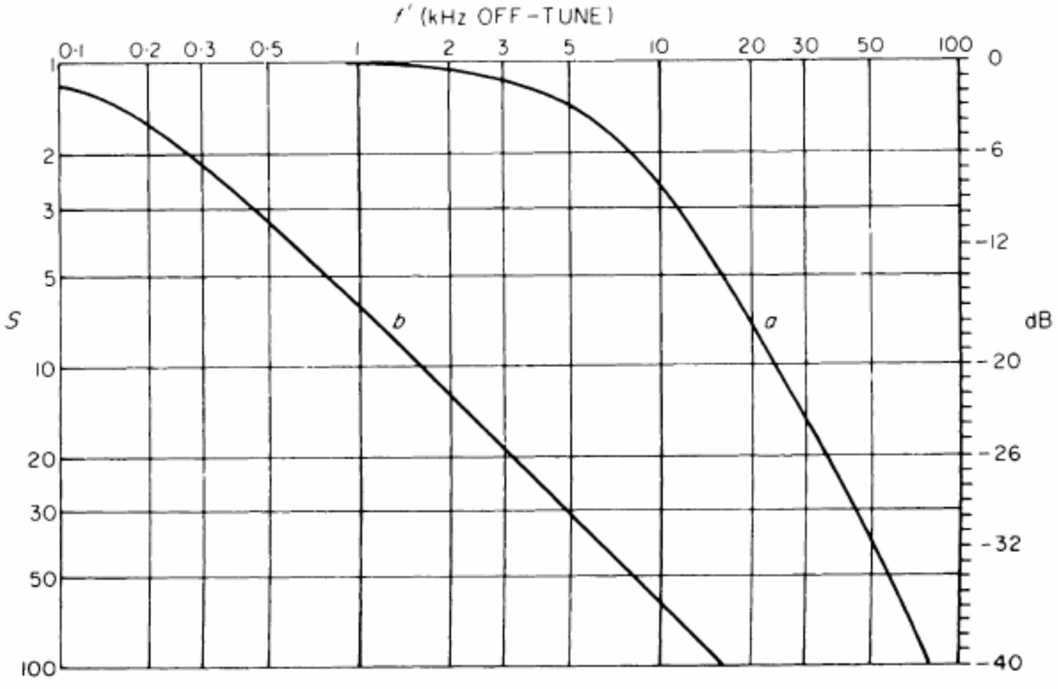


Fig. 20.4 Comparison on an equal-gain basis of a stage of amplification with two tuned circuits (curve (*a*)) with a single tuned circuit given a high magnification by positive feedback (*b*)

it is difficult to avoid enough coupling via the amplifying devices to cause appreciable interaction. Anyway, the use of anything like six r.f. stages, even if technically feasible, would provide far too much

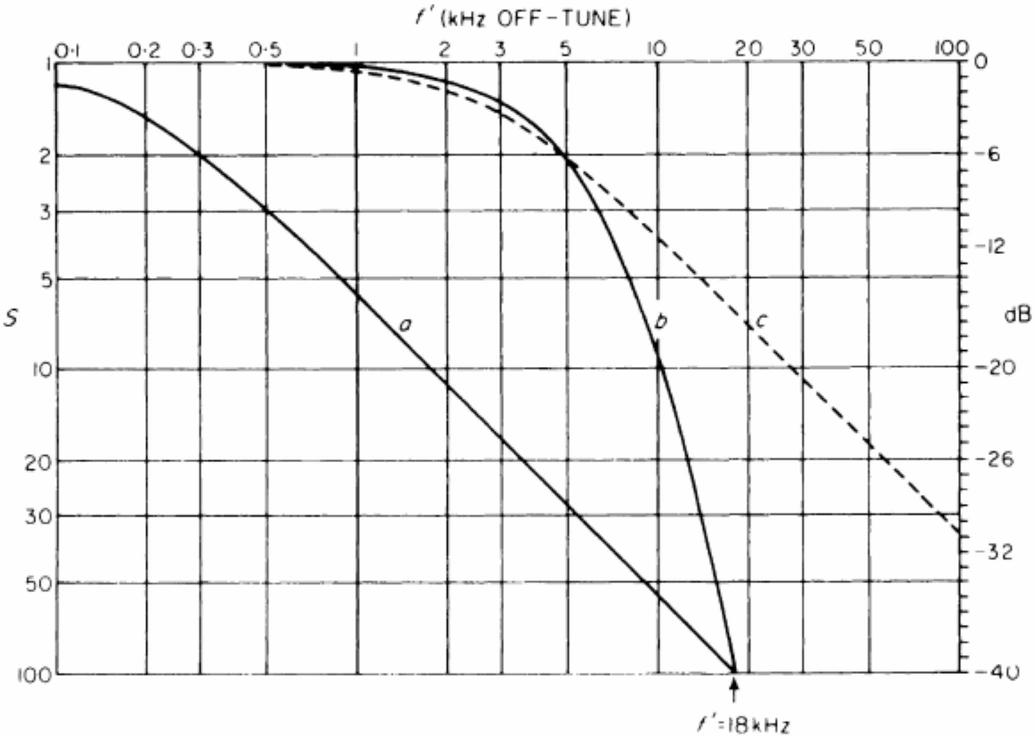


Fig. 20.5 Response curves of one (*a*) and six (*b*) tuned circuits on a basis of equal selectivity at 18kHz, showing how the six preserve sideband frequencies much better. Curve (*c*) refers to a single tuned circuit with its resonance flattened to give equal response at 5kHz off-tune

gain, so most of it would have to be thrown away. But is inter-circuit coupling necessarily bad?

Introducing coupling between tuned circuits makes it more difficult to predict their combined resonance curve. But by a rather more mathematical investigation than is appropriate in a book of this kind it can be shown that beneficial results are obtainable in this way, as we shall now see.

20.6 Coupled Tuned Circuits

There are many ways of coupling two tuned circuits; Fig. 20.6 shows some of them. The feature common to all is a certain amount of shared (or mutual) reactance. In type *a* it is capacitance, C_m; in *b*, self-inductance, L_m; and in *c*, mutual inductance, *M*. Fortunately for calculation, the result does not depend (at any one frequency) on the kind of mutual reactance but only on its amount; so we can denote it simply by *x*. As a matter of fact, method *c* is nearly always used, because no extra component is needed and the coupling can be adjusted simply by varying the spacing of the two coils. Just as L/r is coupling. The question at once arises: with how much? The shape of the curve is different for every value of x. Since the primary purpose of the coupling is to transfer the signal from one circuit to the other, a natural coupling to try for a start is the amount that transfers it most effectively. This is not, as might at first be supposed, the greatest possible coupling. It is actually (with low-loss circuits) quite small, being simply $x = r$.

For this value, called the *critical coupling,* the curve is Fig. 20.7*b*. The curve applies to a pair in which the two circuits have the same L/r; if there is any difference, an approximation can be made by taking the mean L/r. Curve *a,* for uncoupled circuits, is the same as Fig. 20.3 but with all the S figures squared. Although the difference in shape is not enormous, it is quite clear that the curve for the coupled circuits is considerably flatter-topped. But it is also less selective. The advantage gained by the coupling is more obvious if the resistance of the coupled circuits is reduced to half that of the uncoupled circuits, as in Fig. 20.7*c*. If, for example, the L/r of each of the uncoupled circuits represented by *a* were 10, the L/r for each of the coupled circuits represented by *c* would be 20, and the off-tune or sideband frequency treated equally by both combinations (represented by the intersection of *a* and *c*) would be 6 kHz. Although the coupled circuits would thus retain sideband frequencies up to 6 kHz better than the uncoupled circuits they would also give substantially more selectivity.

20.7 Effects of Varying Coupling

Even so, an adjacent-channel selectivity of less than 8 is hardly an adequate projection against interference equal in strength to the wanted signal, still less against relatively strong interference. We may reasonably ask whether any other degree of coupling than critical would give a still better shape; for, if so, any loss in signal transference caused thereby would be well worth while, as it could easily be made up by a little extra amplification. Fig. 20.8 shows what happens when the coupling between two tuned circuits is varied. The curves have been plotted to a linear frequency scale both sides of resonance, to show the kind of shapes seen on an oscilloscope screen (Sec. 23.4).

Reducing the coupling below critical, by making x less than r, not only reduces the signal voltage across the second circuit but also sharpens the peak, so is wholly undesirable. Increasing the coupling splits the peak into two, each giving the same voltage as with the critical coupling at resonance; and the greater the coupling the farther they are apart in frequency. Here at last we have a means by which the selectivity can be increased by reducing the damping of the circuits, while maintaining sideband response at a maximum.

20.8 Practical Tuning Difficulties

It is easy enough to talk about choosing L/r ratios and adjusting couplings to give the best possible response curves. It is not usually very hard to achieve something like the desired results at one frequency. But the demand for receivers permanently tuned to one channel is not large. It is when one attempts to vary the tuning over wide bands of frequency that the troubles really begin. Means must be provided for varying the tuning capacitance or inductance of every circuit continuously and exactly together by one control. Because of the limited frequency band that can be covered in one sweep (Sec. 8.7) switching of inductors is nearly always needed to cover several bands. It would be extremely difficult to maintain all the couplings at exactly the right adjustments and practically impossible to keep L/r always right. When, as is usual, a coil is tuned by a variable capacitor, L of course remains the same over the whole band, but r is much less at the lowest frequency (maximum capacitance) than at the highest. So the selectivity varies accordingly. Even if some way out of this varying selectivity trouble were found, it would only be to run into another. For if L/r is kept constant over a band of frequency, the dynamic resistance R—which as we know is equal to L/Cr (Sec. 8.12)—must be inversely proportional to the tuning capacitance, and therefore directly proportional to the *square* of the frequency. This means very much less amplification at the low frequency end of the band than at the high.

Tuning by varying L instead of C could theoretically avoid both these difficulties at once, if r were made to vary in proportion, for then L/r and L/Cr would both be constant. But this is easier said than done. And we have not begun to think about keeping all this lot stable at all frequencies. So it is not surprising that attempts to produce a really practical high-performance receiver on the lines of variable-frequency r.f. amplification have failed. The generally-accepted solution is the subject of the next chapter.

CHAPTER 21

The Superheterodyne Principle

21.1 A Difficult Problem Solved

Fortunately it is possible to combine the advantages of fixed-tuned selectivity, using a sufficient number of tuning circuits carefully adjusted once and for all in the factory, with the ability to receive over wide ranges of frequency. It is done in the great majority of present-day receivers, under the name of supersonic heterodyne—usually abbreviated to superheterodyne or just 'superhet'. A device called a frequency-changer shifts the carrier wave received from any desired station, together with its sidebands, to whatever fixed frequency has been chosen for the selective circuits. This frequency is called the *intermediate frequency* (i.f.), and the following choices are typical:

For receiving—	*i.f.*
Sound broadcasts (a.m.) on low and medium frequencies	470 kHz
Sound broadcasts (f.m.) on very high frequencies	10.7 MHz
Television broadcasts { vision sound } on v.h.f. and u.h.f.	33.5 MHz 39.5 MHz

The frequencies of all incoming signals in any one of these categories are shifted by the frequency-changer to the same extent. So if, for example, the original carrier-wave frequency of the wanted programme was 865 kHz and it had been shifted to 470 kHz, adjacent-channel carrier waves at 856 and 874 kHz would be passed to the i.f. amplifier as 461 and 479 kHz.

After the desired signal has been amplified at its new frequency and at the same time separated from those on neighbouring frequencies, it is passed into the detector in the usual way, and the remainder of the receiver is in no way different from a 'straight' set.

21.2 The Frequency-Changer

The most important thing to grasp in this chapter is the principle on which the frequency-changer works. All else is mere detail.

Suppose a sample of the carrier wave coming from the desired station during one hundred-thousandth of a second to be represented by Fig. 21.1*a*. As it contains 10 cycles, the frequency cycles being so numerous that they are hardly distinguishable. And this time it is fully modulated; the sample is sufficient to show two whole cycles of modulation frequency. The local oscillation, which is certain to be much stronger, and of course is unmodulated, is shown at *b*. When *a* is added to it (*c*), nothing happens to it at the troughs of modulation, where *a* is zero, but elsewhere the cycles of *a* alternately strengthen and weaken those of *b*, causing its amplitude to vary at the beat frequency, in the manner shown. The amplitude of beating is proportional to the amplitude of *a*, and the frequency of beating is equal to the difference between the frequencies of *a* and *b*. The next picture, *d*, shows the result of rectifying *c*. We now have, in addition to other frequencies, an actual signal at the beat frequency (or i.f.), shown at *e* after it has been separated from the rest by the i.f. tuned circuits. The result of rectifying *e* in the detector is *f*; and after the remains of the i.f. have been removed by the detector filter the final result (*g*) has the same frequency and waveform as that used to modulate the original carrier wave.

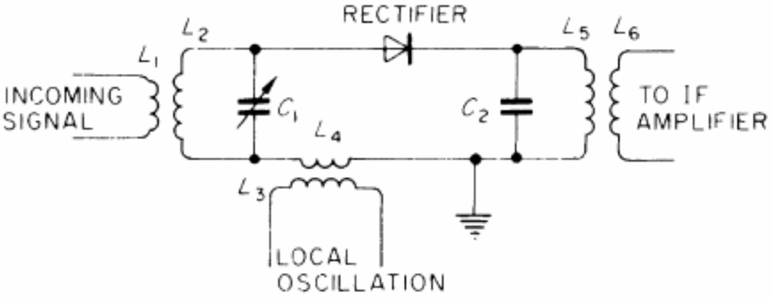


Fig. 21.3 Simple type of frequency-changer, in which the incoming signal and the local oscillation are added together by connecting their sources in series, and then rectified

Fig. 21.3 shows an elementary way in which the process just described can be carried out. The incoming signal and the local oscillation are added together by connecting their sources in series with one another, and the two together are fed into a rectifier—some kind of diode. C_1 is used for tuning L_2 to the signal, so as to develop the greatest possible voltage at that frequency. There is, of course, no difficulty in getting an ample voltage from the local oscillator. C_2 serves the double purpose of tuning the transformer primary (L_5) to the intermediate frequency, and of providing a short-circuit for the two input frequencies, which are not wanted there.

21.3 Frequency-Changers as Modulators

If the rectifier in Fig. 21.3 had been omitted, or replaced by an ordinary linear resistor, no new signal frequency would have been created. In linear circuits, currents of different frequencies can flow there is a frequency equal to that of *b plus* that of *a* is not at all obvious, but its existence can be shown mathematically, or by actual experiment—picking it out from the rest by tuning a sharply resonant circuit to it.

The purpose of the local oscillator can now be seen as a means of varying the conductance of a part of a circuit carrying the incoming signal, amplitude-modulating it and thereby creating signals of at least two new frequencies.

Either of these could be chosen as the i.f. in a superhet, but the lower or difference frequency is the one usually favoured, because it is easier to get the desired amplification and selectivity at a relatively low frequency. See also Sec. 21.6 for a more important reason.

21.4 Types of Frequency-Changer

A frequency-changer therefore consists of two parts: an oscillator, and a device whereby the oscillation modulates or varies the amplitude of the incoming signal. This device is commonly called the mixer. If the word 'mix' is understood in the sense a chemist (or even an audio engineer) would give it—just adding the two things together—then it is a very misleading word to use. Simply adding voltages or currents of different frequencies together creates no new frequencies, just as merely mixing oxygen and hydrogen together creates no new chemical substance. So from this point of view 'combiner' would be a better name for the second part of a frequency-changer. It can also be fairly termed the modulator.

The oscillator presents no great difficulty; almost any of the circuits shown in Chapter 14 could be used. It is, of course, necessary to provide for varying the frequency over a suitable range, in order to make the difference between it and the frequency of the wanted station always equal to the fixed i.f.

The simplest type of modulator or 'mixer' is the plain non-linear element—a diode—already seen in Fig. 21.3. Frequency-changers essentially of this type are used in radar and other receivers working on frequencies higher than about 800 MHz, for reasons which will appear shortly. But they have certain disadvantages which have resulted in their having been superseded for most other purposes. For one thing, adjustment of C_1 is liable to affect the frequency of the oscillator. The local oscillations on their part are liable to find their way via L_2 and L_1 to the antenna, causing interference with other receivers. In short, the signal and oscillator circuits interact, elsewhere than in the proper direction—through the rectifier. And the rectifier does not amplify; on the contrary, it and its output circuit tend to damp the signal presented to it by the tuned circuit L_2C_1.

The principle of a commonly used type of frequency changer is multiplied, producing, at the output of the transistor, components at the sum and difference of the two input frequencies. The difference component is selected by the tuned transformer in the drain circuit and is passed on to the i.f. amplifier for further amplification.

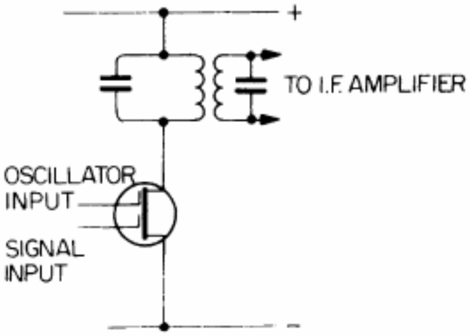


Fig. 21.6 One of several types of device which can be used for frequency-changing—a dual-insulated-gate f.e.t.

21.5 Conversion Conductance

The effectiveness of an ordinary amplifying or oscillating type of transistor is, or can be, expressed as the mutual conductance, in milliamps of signal current in the output circuit per volt of signal applied at the input. In a frequency-changer, however, the current at signal frequency is a waste product, to be rejected as soon as possible; what we are interested in is the milliamps of i.f. current per volt of signal-frequency. This is known as the *conversion conductance,* denoted by g_c. Obviously it does not depend on the device alone but also on the oscillation voltage. If this is too small, it does not vary the mutual conductance over its full range; consequently the signal current is modulated to only a part of its full amplitude, and the i.f. current is less than it need be. The conversion conductance reaches its maximum when the local osillation is enough to bring the mutual conductance to zero during its negative half-cycles and to maximum during as much as possible of its positive half-cycles. The positive part of Fig. 21.2*c* is therefore reproduced with the full g_m, while the negative part is completely suppressed; the net result is the same as with a perfect rectifier, Fig. 21.2*d*. The incoming signal, however weak, is swept right round the bend by the oscillation. That explains why a superhet is sensitive to the feeblest signals, which, if applied straight to the detector, would make no perceptible impression.

In Fig. 21.7*a*, a small sample of oscillation plus signal (such as Fig. 21.2*c* or 21.1*c*) is shown in relationship to the input voltage scale of an active device curve. (The fact that the oscillation and signal may actually be applied to different terminals does not affect the essential principle.) Since the i.f. fluctuations are due to the signal alternately strengthening and weakening the local oscillation, their amplitude is, as we have seen, equal to that of the signal. Neglecting the effect

21.6 Ganging the Oscillator

We have seen that the intermediate frequency is equal to the difference between the signal frequency and the oscillator frequency. With an i.f. of 470 kHz, for example, the oscillator must be tuned to a frequency either 470 kHz greater or 470 kHz less than the signal.

Suppose the set is to tune from 525 to 1605 kHz. Then, if higher in frequency, the oscillator must run from (525 + 470) to (1605 + 470), i.e., from 995 to 2075 kHz. If, on the other hand, the oscillator is lower in frequency than the signal, it must run from (525–470) to (1605–470), or 55 to 1135 kHz. The former range gives 2.09 as the ratio between highest and lowest frequency; the latter 20.6. Since even the signal-circuit range of 3.06 is often quite difficult to achieve, owing to the irreducible capacitance likely to be present in a finished set, the oscillator range from 995 to 2075 kHz is the only practical choice.

The next problem is to control both signal-frequency circuit tuning and oscillator frequency with one knob. Because there must be a constant frequency difference between the two, it hardly seems that what is called a ganged variable capacitor—consisting, in effect, of two identical capacitors on the same spindle—would do. As we have just seen, the ratio of maximum to minimum capacitance is quite different. This difference, however, can easily be adjusted by putting a fixed capacitance either in parallel with the oscillator capacitor to increase the minimum capacitance, or in series to reduce the maximum. The ratio of maximum to minimum having been corrected in either of these ways, correct choice of inductance for the oscillator coil will ensure that it tunes to the right frequency at the two ends of the tuning-scale.

In the middle of the tuning range, however, there will be an error. The oscillator tuning capacitance will be too large or too small, depending on whether a parallel or series capacitor is used to restrict the oscillator range. It is found that a judicious combination of the two methods, using a small parallel capacitor to increase the minimum a little, and a large series capacitor to decrease the maximum a little, will produce almost perfect 'tracking' over the whole waveband.

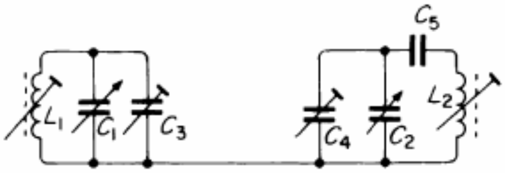


Fig. 21.8 Preselector and oscillator tuning circuits arranged to maintain a constant frequency difference

Fig. 21.8 shows these tuning arrangements applied to one frequency band. $L_1C_1C_3$ is a preselector tuning circuit; there is often at least one other, either at the input and output of a stage of

21.7 Whistles

A superheterodyne is liable to certain types of interference from which an ordinary set is free. The most noticeable result of these is a whistle, which changes in pitch as the tuning control is rotated, as when using a 'straight' receiver in an oscillating condition, but generally less severe. There are many possible causes, some of which are quite hard to trace.

The best-known is *second-channel* or *image* interference. In the preceding section we have seen that it is usual for the oscillator to be adjusted to a frequency higher than that of the incoming signal. But if, while reception is being obtained in this way, another signal comes in on a frequency higher than that of the oscillator by an equal amount, an intermediate frequency is produced by it as well. If the frequency difference is not *exactly* the same, but differs by perhaps 1 kHz, then two i.f. signals are produced, differing by 1 kHz, and the detector combines them to give a continuous note of 1 kHz.

An example will make this clear: Suppose the signal desired works on a frequency of 200 kHz, and the i.f. is 470 kHz. When the receiver is tuned to 200 kHz the oscillator is 470 kHz higher, 670 kHz, and yields a difference signal of 470 kHz, to which the i.f. amplifier responds. So far all is well. But suppose now a signal of 1137 kHz is also able to reach the frequency changer. It will combine with the oscillation to produce a difference signal of 1137–670, = 467 kHz. This is too near 470 kHz for the i.f. amplifier to reject, so both are amplified together by it and are presented to the detector, which produces a difference frequency, 3 kHz, heard as a high-pitched whistle. Slightly altering the tuning control alters the pitch of the note, as shown in the table:

Set tuned to:	*Oscillator at:*	*I.F. signal due to 200 kHz signal (wanted)*	*I.F. signal due to 1,137 kHz signal (interfering)*	*Difference (pitch of whistle)*
kHz	kHz	kHz	kHz	kHz
195	665	465	472	7
196	666	466	471	5
197	667	467	470	3
198	668	468	469	1
198½	668½	468½	468½	0
199	669	469	468	1
200	670	470	467	3
201	671	471	466	5
202	672	472	465	7
203	673	473	464	9

An example of how second-channel interference is produced

CHAPTER 22

Radio-frequency and Intermediate-frequency Amplification

22.1 Amplification So Far

We have examined methods of audio-frequency amplification at some length in Chapter 19. Broadly, they consist of amplifying devices—transistors—connected together by couplings which are sometimes called *aperiodic* because they are designed to have no natural period or resonance but to treat all frequencies alike, at least those within the audio band. Since then we have assumed the practicability of amplifiers that respond only to the very limited band of frequencies included in one sound channel or the much wider (but, in proportion to the carrier frequency, still small) band in a television channel. In these the inter-stage couplings consist of tuned circuits. For an i.f. amplifier these couplings are preset, whereas for amplification at the original signal frequency they have to be varied by the user if he wants to receive on more than one channel. Although 'r.f.' includes i.f., where both are mentioned in the same context it can be taken that it means the original incoming signal frequency.

22.2 Active-device Interelectrode Capacitances

The electrodes of active devices, together with their connecting leads, act as tiny capacitors, with effects on the behaviour of the circuit as a whole. These capacitances are normally only a few picofarads and, to get some idea of what they mean in practice let us calculate the reactance of 4 pF at various frequencies (Sec. 6.3). The reactance varies from 4 MΩ to 40 Ω and suggests that these small capacitances have negligible effect on the operation of active devices at audio frequencies. But at much higher frequencies the reactance is low enough to play a vital part. This is illustrated in the v.h.f.

Frequency	*Reactance*
10 kHz	4 MΩ
100 kHz	400 kΩ
1 MHz	40 kΩ
10 MHz	4 kΩ
100 MHz	400 Ω
1 GHz	40 Ω

We were then considering only resistances (or conductances) and only plus and minus, so that the only effects were increases or decreases in value. In this chapter, however, we have to take into account reactances as well as resistances, and consequently all possible phase relationships as well as just 0° and 180°.

22.3 Miller Effect

To take a simple case first, consider Fig. 22.2*a*, in which C is a capacitor hidden in our now well-worn black box, in series with a special sort of meter that measures the charge passing into the capacitor when the switch is closed. As capacitance is equal to the charge required to raise the p.d. across it by 1 V (Sec. 3.2) we have here a means of measuring C. But suppose (b) that when we connect our 1 V to one end of C a battery giving A volts is applied at the other end with polarity as shown. The charging voltage is $A+1$, so the charge is $A+1$ times as much as that due to 1 V only. Knowing nothing of the hidden battery, we would suppose the box contained a capacitance equal to $(A+1)C$, as at c. So far as external electrical tests can tell, the boxes b and c are identical.

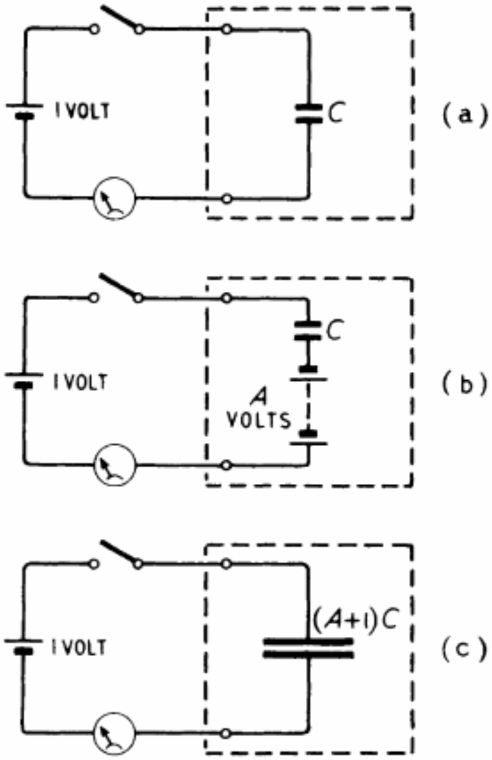


Fig. 22.2 Showing how the introduction of an extra voltage, such as that fed back from output to input in an active device, can be equivalent to a multiplied capacitance

This is exactly what happens in an amplifier having a voltage amplification A, if the output load is purely resistive at the frequency concerned—it could be the dynamic resistance of a tuned circuit. As A might be 50 or more, an internal input/output capacitance of only

2 pF could have the same effect as one of 100 pF. This would be serious enough, but at a slightly different frequency or output tuning the phase angle could shift nearly 90° one way or the other. The effect can be considered by analysing the feedback into two parts, one exactly negative (180°) and the other 90° different. The 180° component magnifies the internal capacitance as explained. In so far as the output load was capacitive, however, the current fed back to the input terminal would be in such a phase as to appear to input voltages as conductance, which would have a loading or damping effect. An inductive output load, on the other hand, would feed back current in the opposite phase, giving the effect of negative conductance. If this was enough to offset the loss conductance of the input circuit, oscillation would result.

These feedback effects via the internal input/output capacitance of an active device are named after their discoverer, J. M. Miller. With tuned amplifier couplings, Miller effect causes at least tuning difficulties and can even cause oscillation. Now we can see that if resistance coupling were used at radio frequencies the inevitable stray capacitance across it would lead to a capacitive component of feedback current tending to increase the input conductance—which one tends to assume is practically zero with an f.e.t. So Miller effects are, in general, bad.

22.4 High-frequency Effects in Transistors

There are two other causes of apparent input conductance, which with normal design do not seriously impair the usefulness of active devices as amplifiers at frequencies below about 50 MHz, but need to be considered at very high frequencies. One is inductance of the emitter or source lead (Fig. 22.3). This is common to both input and output circuits and produces the effect of input conductance by feedback.

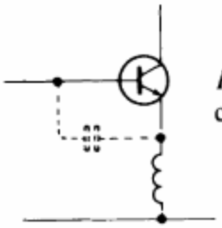


Fig. 22.3 Feedback due to emitter lead inductance and base-emitter capacitance of a transistor

The other is the time taken for charge carriers to travel between emitter and collector in a bipolar transistor—known as *transit time.* Besides the ordinary capacitive input current due to the input voltage across c_{be}, an additional current is induced by the movements of charge carriers across the base region. A small fraction of these is shown for an r.f. transistor under working conditions. Another advantage is that these parameters are fairly simply related to the complex *h* values. Over a small band of frequency such as the 470 kHz i.f. band the hybrid π circuit of Fig. 22.4 can be replaced by a slightly simpler equivalent in which $r_{bb'}$ is eliminated by allowing for it in the values of the other components. However, to see how the more comprehensive circuit works in practice let us draw phasor diagrams for it at two widely different frequencies.

22.5 Transistor Phasor Diagrams

Fig. 22.5 shows the Fig. 22.4 equivalent circuit fitted with a resistive load of 20 kΩ and labelled for phasors. Let us first draw diagrams for a frequency low enough for the capacitances to be neglected. As usual, the best starting place is the output, and to get realistic proportions let us calculate actual values, assuming an output *ec* of 1 V.

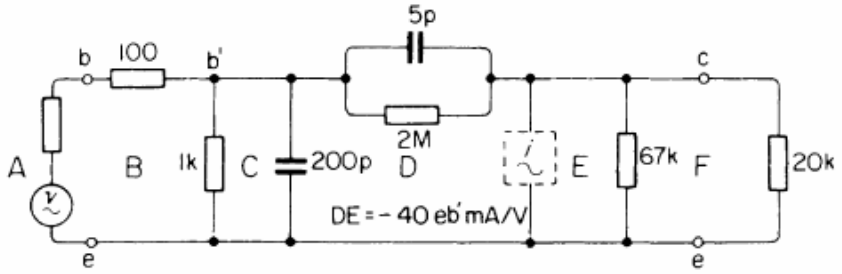


Fig. 22.5 Hybrid π circuit marked with typical values and phasor-diagram labels

Because the input voltage will be relatively small, we choose a scale that brings only 1/400th of *ec* into the diagram, Fig. 22.6. We calculate and plot *AF, FE* and *DA*, which are 50, 15 and 0.5 μA respectively. For this purpose $r_{b'e}$ is negligible compared with $r_{b'c}$. So the current generator has to supply 65.5 μA, and making use of the generator equation given we find *eb'* to be 65.5/40 = 1.64 mV. *DC* is zero in this case, so *DB* is 1.64/1 = 1.64 μA. That completes the current diagram and reveals *AB* as 0.5 + 1.64 = 2.14 μA, from which *bb'* follows as 0.214 mV, to complete the voltage diagram. The input voltage, *eb*, is 1.64 + 0.214 = 1.854 mV, so the voltage gain (*ec/eb*) is 1 000/1.854 = 540, the current gain is 23.4, and the power gain 12 600, or 41 dB. Note that $r_{b'c}$ causes some reduction in gain by negative feedback (*BD* < *BA*). The input resistance, *eb/AB,* is 866 Ω.

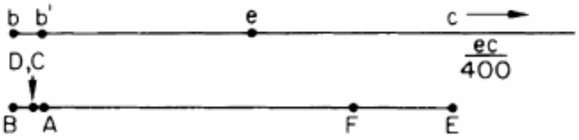


Fig. 22.6 Low-frequency phasor diagram for Fig. 22.5

Let us now see what happens at 470 kHz. At this frequency the reactance of $c_{b'e}$ is 1.7 kΩ. And $c_{b'c}$ is 68 kΩ, which in parallel with $r_{b'c}$ is very nearly equivalent to the same capacitance in series with 2.3 kΩ of resistance. The total impedance (Sec. 6.8) is hardly more than 68 kΩ. We begin as before with 1 V of *ec*, and assuming the same load we get the same *AF* and *FE* (Fig. 22.7). But *DA* is very different both in magnitude and phase angle. The feedback impedance is still large enought for $r_{b'e}$ to be neglected in comparison with it, but is much less than before, so *DA* is 1000/68 = 14.7 μA; and by drawing the impedance triangle (or by calculation) we find that it leads *ec* by 88°. Plotting it fixes *D* and gives us *ED* as 67.5 μA, the total generator current. Using the equation again we find *eb′* to be 1.7 mV, so *DC* is 1 μA leading it by 90°. *CB* is 1.7 μA, in phase, making *BA* 16 μA and *bb′* 1.6 mV. The completed diagram tells us that *eb* is 2.7 mV, so the voltage gain is down to 370 and the current gain more drastically down to 3.1. The power gain is not simply A_vA_i this time, because there is a phase difference of 37° between input voltage and current. The input power is therefore not ***eb.AB*** but ***eb.AB*** cos 37° (Sec. 6.9), so the power gain is $A_vI_v/\cos 37°$, = 1450, or 31.6 dB. The ratio ***eb/AB*** indicates an input impedance of 170 Ω, which cannot be accounted for by the values of $r_{bb'}$, $r_{b'e}$ and $c_{b'e}$ alone, but clearly points to Miller effect. The apparent value of $c_{b'e}$ is in fact 3.2 nF; 16 times its real value.

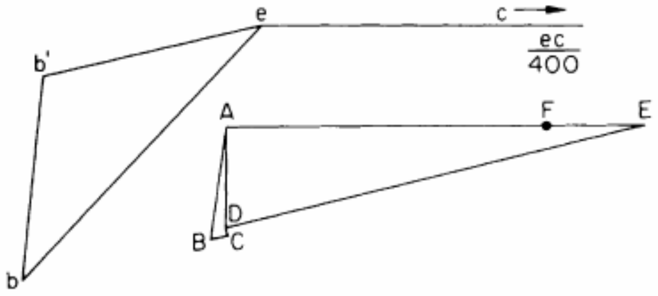


Fig. 22.7 470 kHz phasor diagram for Fig. 22.5

Nevertheless this transistor is capable of useful i.f. gain, and in a suitable circuit could go considerably higher. The data we have gathered by drawing the phasor diagrams will help us soon in designing an i.f. amplifier.

22.6 Limiting Frequencies

The very important transistor parameter that we have so often been using in our studies at low frequencies is the short-circuited-output current gain, h_f, or h_{fe} when applied to the common-emitter mode in symbol for it is f_T, but this one is sometimes taken to mean the product of low-frequency h_{fe} and bandwidth. As the bandwidth is regarded as the frequency band up to the −3 dB point, this interpretation of f_T gives it the same value: 40×0.8 MHz in this example. But this is only so in this rather over-simplified theory of the relationship between h_{fe} and frequency, and in practice the different definitions of f_1 and f_T may result in slightly different values.

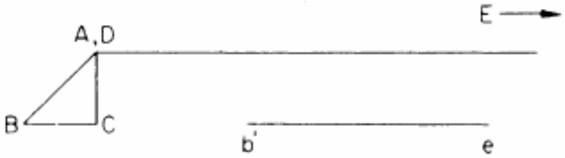


Fig. 22.10 Simplified high-frequency phasor diagram for a transistor, derived from Fig. 22.9

The same zero-dB frequency is also normally nearly the same as that at which h_{fb} is 3 dB down. It therefore is a measure of common-base bandwidth. And as the low-frequency h_{fb} is almost equal to 1, the gain × bandwidth defintion of f_T leads to about the same f_T for common-base as for common-emitter. Although therefore a transistor in common-base gives no current gain and so less voltage and power gain than in common-emitter, what it has it maintains up to a much higher frequency: about h_{fe} times higher, in fact.

22.7 An I.F. Amplifier

For this first attempt at designing an i.f. amplifier we take single tuned circuit couplings, such as are used in most portable transistor receivers. We start from the bandwidth, f_B, which we found in Sec. 20.4 is equal to the resonant frequency (470 kHz in this case) divided by the Q of the tuned circuit. A practical bandwidth is 4 kHz each side of resonance, or 8 kHz altogether. So the required Q is 470/8; say 60. This, it should be understood, is the Q under working conditions, loaded by the output resistance of the preceding transistor and the input resistance of the following one. At the moment we do not have either of these data for the transistor represented by Fig. 22.5. If we can assume that the load is 20 kΩ resistance we have the input *impedance* at 470 kHz, 170 Ω, but Fig. 22.7 shows that the input current leads the input voltage by 37°. A little calculation on the lines of Chapter 6 shows that this is equivalent to 216 Ω in parallel with 1.23 nF. The output resistance is more difficult. Another phasor diagram is needed to derive it from Fig. 22.5, and we would have to know the signal source impedance as well. If we assume it is made equal to the input resistance, the output resistance comes out at about 25 kΩ, in parallel with capacitance.

Although the tuning product LC is fixed for us by the resonance frequency (Sec. 8.7) we are free to choose the ratio L/C. If C is small, so that most of it is made up of transistor and wiring capacitances, it dimensions, these are fixed, normally at about 470 kHz or 10.7 MHz. Ceramic filters combine good i.f. tuning with small size.

Another type of filter now used in i.f. amplifiers is the surface-acoustic-wave (s.a.w.) device. By means of a transducer* the signal is transmitted along the surface of a piezo-electric material in the form of minute deformations of the surface. By shaping the surface suitably the effect of a bandpass (or other type of) filter can be achieved and the signal can be recovered by use of another transducer. Extremely small robust filters with stable characteristics can be produced by this means.

22.8 R.F. Amplification

Amplification at the original signal frequency, if that is lower than about 30 MHz, is now seldom used except in high-performance 'communications' receivers; but when it is used the techniques are similar to those for i.f., with the addition of variable tuning. This is done by one or more extra sections on the ganged tuning capacitor. Because the i.f. amplifier is relied upon for most of the selectivity, the preselector tuning circuits do not have to be very sharp, so slight errors in ganging are not serious. So long as the tuning capacitor sections are reasonably alike, all that is needed is to make the inductances and minimum capacitances equal. The only difficulty arises in the first tuned circuit, to which the antenna is coupled; this problem is considered in Sec. 22.12.

Preselector (r.f.) amplifiers are most used at frequencies higher than about 30 MHz, because noise generated in the frequency changer tends to be greater the higher the frequency, and signals tend to be weaker. So the all-important signal/noise ratio is usually not good enough without some amplification at the original frequency. This is rather awkward, because the difficulties in achieving such amplification increase with frequency. And of course only low-noise devices can be considered.

Feedback via the internal base/collector capacitance of transistors makes the common-emitter tuned amplifier useless at v.h.f. or u.h.f. But the common-base circuit is suitable, and to understand this we must know a little about the principles of screening.

Feedback occurring via a capacitance can obviously be avoided by getting rid of the capacitance. The capacitance between any two objects can be eliminated by placing between them as a screen an earthed metal sheet of sufficient size. The operation of such a screen can be understood by considering Fig. 22.13*a*, in which a stray so tuned circuits in the form of quarter-wave 'lines' as described in Sec. 16.10 are used.

---

* A transducer is, in general, a device able to convert an input signal in one form of energy into a corresponding signal in another form. Thus a microphone is an acoustic/electrical transducer. The transducers in the s.a.w. filter are electrical/mechanical.

Fig. 22.16 shows a typical arrangement, in which L_1 is the input tuned circuit consisting of a quarter-wave rod which is the inner conductor and the surrounding metal screen (shown dashed) is the outer conductor. It is inductively coupled to the u.h.f. antenna feeder on one side and to the emitter of the common-base transistor on the other. The collector is connected to a tapping on the output tuned circuit L_2. The d.c. is fed through a choke coil that prevents the signal voltage from being short-circuited. Note that all the feeds pass through the screen via 1 nF 'feed-through' capacitors which effectively earth all signals to the screen. L_2 is loosely coupled to L_3 through a hole in the screen, so that the pair are a band-pass filter represented in conventional symbols by Fig. 20.6*c*. Its output is inductively coupled to the common-base frequency changer. L_4 is the oscillator tuning circuit, back-coupled to the emitter via C_2. The u.h.f. and i.f. signals in the collector circuit are separated by a choke coil, and the i.f. signals are developed across L_5. All four line circuits are tuned by a four-section ganged capacitor $C_{1a\text{-}d}$.

In modern tuners the capacitors C_{1a} to C_{1d} are likely to be replaced by varactors controlled from a single voltage as described in Sec. 21.6.

22.9 Screening

The basic principle of screening was explained in the previous section with particular reference to active device electrodes; the same principle applies to the external screening of circuits. The object is to prevent capacitive coupling between two circuit elements by interposing a conducting sheet which is either directly earthed or has such a large capacitance to earth that it is effectively at zero signal potential. Metal sheet or foil is the usual screening material, but its effectiveness is not greatly reduced by perforations, and for some large-scale purposes wire netting is good enough. Because even very small stray capacitances have sufficiently low impedance at very high frequencies to form undesired circuit paths, the need for screening generally increases with frequency. The complete screening of the u.h.f. circuits in Fig. 22.16 is an example of this.

One cannot always assume that because two points are connected to such a screen they must both be at the same signal potential. If a substantial signal current flows between points some distance apart there may well be enough impedance between them to give rise to an appreciable signal p.d. and therefore a risk of undesired coupling. So care has to be taken in laying out u.h.f. circuits.

Signal leakage via d.c. or relatively low-frequency leads is another possibility to watch. Fig. 22.16 includes two examples of u.h.f.

vertical slot parallel to the coil axis (Fig. 22.17) would prevent magnetic screening whereas a horizontal slot would hardly affect it.

So it is possible to screen both kinds of field, or either of them alone.

In general the permeability screen is the more effective magnetic type at low frequencies, including z.f., and the conductive screen at high frequencies. The permeability type increases the inductance of the enclosed inductor, especially if it is shaped so as not to be a good conductive screen. Provided that this is not objectionable, the screen need not be widely spaced from the inductor. A closed magnetic core acts as a magnetic screen by short-circuiting the magnetic flux. A conductive screen reduces the inductance and tends also to increase its resistance, the closer it is to the inductor. So the inductor itself should be designed to have as little external field to screen as possible.

22.10 Cross-Modulation

The emphasis on distortion in the chapter on a.f. amplification has been notably lacking in the present chapter. The reason is that so long as the shape of the modulation waveform is preserved, the shapes of the individual r.f. cycles that carry the modulation are unimportant. They could all be changed from sine waves to square waves or any other shape, and indeed this is actually done in f.m. receivers by the limiter, which chops all the cycles down to a constant size, regardless of distortion. That would not be allowable in an a.m. receiver, of course (the purpose of the limiter being in fact to remove any a.m.) but so long as the relative sizes of the cycles were preserved they would still define the original modulation envelope. A perfect detector cuts off half of every cycle, yet this drastic r.f. distortion introduces no distortion of the modulation. It is the imperfect or non-linear detector (Sec. 18.4) that distorts the modulation by altering the relative proportions of the r.f. cycles. But unless the modulation is fairly deep it is surprising how little it is distorted by even a considerable curvature in the r.f. amplifier's output/input characteristic. It must be remembered that except for the output of the stage feeding the detector the amplitude of the r.f. signals is so small that it is unlikely to experience appreciable non-linearity.

There is an important exception to this. If signals very much stronger than those to which the receiver is tuned are not very widely different in frequency they may get some way into the receiver before being tuned out by its selectivity. In a superhet most of the selectivity is concentrated in the i.f. amplifier, for the reasons given in Sec. 20.8. So such signals may well be present in the first stage at considerable amplitude. What happens there is very similar to the frequency-changer action illustrated in Fig. 21.4*b*, the interfering interference. Such interference is unlikely on the v.h.f. and u.h.f. bands, so the problem seldom intrudes there, which accounts for the fact that sometimes there is no variable tuning at all before the first stage (Figs. 22.14 and 15, for example), but broadcast receivers for the medium frequency band (especially if used with large antennas) and communications and other specialist receivers must be designed with cross-modulation in mind.

It is fortunate that when a carrier wave is tuned in it affects the detector in such a way that if some interfering modulated carrier wave is present its modulation is rendered more or less ineffective, i.e., the carrier is truly demodulated. In this way the selectivity of the receiver is apparently increased. With sharp tuning, the percentage modulation of the wanted a.m. signal is also reduced, reducing detector distortion. So it may pay to tune sharply in the interests of selectivity and detector linearity, and make up for the resulting weakening of outer sidebands by upper frequency compensation in the a.f. department.

22.11 Automatic Gain Control

Receivers for sound are usually designed for a certain signal amplitude at the detector, such that at a moderate depth of modulation (say 30%) there is enough a.f. gain to give maximum output, plus a bit in reserve. The volume control, to enable the output to be adjusted between more-than-maximum and zero, is usually placed between the detector and the first a.f. stage. Ideally, all carrier waves to which the receiver is tuned would arrive at the detector at the same amplitude. This implies that the r.f. gain is always adjusted inversely to the strength of the signal as it comes from the antenna. This adjustment is not one that the user of the receiver wants to be bothered to make, so most receivers are designed to do it automatically, by means of what is called *automatic gain control* (a.g.c.). To get close to the ideal some fairly complicated circuitry is needed. Here we shall look at only the simplest methods.

The general principle is to use the rectified carrier wave voltage part of the detector output (the d.v.) to alter the bias of one or more of the preceding stages in such a way as to reduce their gain. The choice of stages calls for careful consideration. Control of preselector stages, if any, is most desirable, but the process usually increases non-linearity so the risk of cross-modulation must be watched. The type of frequency changer that combines oscillator and 'mixer' is not available for a.g.c., because oscillation might be stopped thereby, or at least mistuned. The last i.f. stage is handling signals at their maximum strength, so must not be allowed to become non-linear enough to distort the modulation. In most transistor receivers it is customary to control only the first i.f. stage.

This feature is called—not very aptly, for time is not involved—*delayed* a.g.c.

Note that the a.g.c. potential divider system in Fig. 22.19 applies a small forward bias to the detector diode, sufficient to bring the working point to the sharpest point on its characteristic curve (see Fig. 9.14) and thereby increase its response to weak signals.

Oddly enough, a.g.c. is possible with transistors by applying control bias of the opposite polarity to that just indicated. For this forward a.g.c. (as it is called) to work, there must be a suitable amount of resistance in the collector circuit. So when a.g.c. bias is applied and increases the collector current, the voltage reaching the collector falls and as the transistor begins to run into saturation its gain falls fairly steeply. At the same time its output resistance falls and damps the tuned circuit into which it works, increasing the bandwidth and sideband acceptance. The associated loss of selectivity can be accepted because it takes place only when a comparatively strong carrier wave is tuned in and the gain is low, so interference is unlikely. Special r.f. transistors are required to obtain satisfactory results from forward control. The characteristics of reverse-control r.f. transistors are not suitable.

A.g.c. is hardly necessary for f.m., because the limiter irons out the differences in strength between incoming signals. In the exceptional cases where signals of such widely different strengths are encountered that the limiter would have difficulty in coping with the full range, the nearly constant d.c. output of the f.m. detector would obviously be ineffective for a.g.c. purposes but the long-time-constant circuit of a ratio detector (Sec. 18.16) provides a convenient source of a.g.c. voltage.

In the vision section of a television receiver the provision of a.g.c. is complicated by the fact that the vision signal has no constant mean carrier amplitude; it varies with the average light and shade in the picture. So if a.g.c. is applied in the same way as for sound, it tends to brighten scenes that are supposed to be dark and darken scenes that are supposed to be bright. In designs where trouble is taken to exclude this defect, certain parts of the composite vision signal that are not reproduced on the screen, and which represent the strength of the incoming signal, are selected by what is called a gating circuit for use as a.g.c. bias.

22.12 The Antenna Coupling

The receiving antenna can be regarded as a signal generator and the receiver as the load, and the signal voltage is fixed by the incoming field strength and the nature of the antenna, so the maximum power law (Sec. 11.8) does apply. But the transfer of power has to be compatible with selectivity and easy tuning.

from that of the secondary and so the less the capacitance that is transferred into it. The greater signal loss is compensated by the stronger signal the larger antenna brings in. Signal transfer tends to be poor at the highest frequencies, at which L_2C_2 is most out of tune with the antenna, so a small capacitor C_3 is sometimes included to help things along. Another method (Fig. 22.23*b*) has the merit of simplicity and of ensuring that whatever antenna is used the capacitance it adds to the tuned circuit must be less than C_4. And the range of signal strength brought in by antennas of different sizes is to some extent levelled out, for the larger it is the more its coupling is reduced by the capacitance C_4 (which is usually only a few pF) being so much less than its own. At the same time the effects of diverse antenna resistances are reduced.

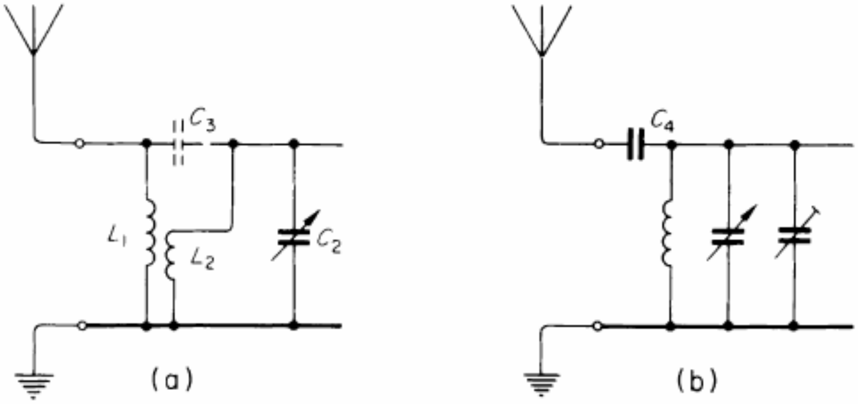


Fig. 22.23 Two other methods of antenna coupling: (*a*) a relatively high-inductance primary coil (L_1) augmented if necessary by top capacitance (C_3); (*b*) small series capacitance

Providing for v.h.f. dipoles is comparatively easy because their resistance is specified, their reactance is fairly small or even zero, and the frequency ratio is usually quite limited.

Portable receivers with internal antennas that are in effect large tuning coils are easier still, for the designer has everything under his control.

22.13 Klystrons and Travelling-wave Tubes

We saw in Sec. 22.4 that the time taken by charge carriers to cross the base region of a transistor limits the upper frequency at which it can be successfully used as an amplifier. But, as mentioned in that section, there is a class of r.f. amplifying devices which relies on transit time for its operation. We have already encountered one of these devices; it is the klystron described in Sec. 14.14. These devices need a significant transit time because, during it, the electron beam is gathered into bunches which generate the output of the tube.

Another microwave device depending on transit time is the travelling-wave-tube (t.w.t.). This uses the same principle, i.e., the electron beam is bunched by the electric field set up by the input

CHAPTER 23

Cathode-Ray Tubes: Television and Radar

23.1 Description of Cathode-Ray Tube

The preceding chapters have covered the general principles underlying all radio systems (and incidentally quite a number of other things such as deaf aids and r.f. cookers), together with sufficient examples of their application to provide at least the outlines of a complete picture of the reception of sound programmes. There remains one device which is so important for such purposes as television and radar that this book would be incomplete without a chapter on it; the cathode-ray tube which, when used to display television images, is known as a picture tube.

In many respects it resembles a valve; the differences arise from the different use made of the stream of electrons attracted by the anode. Both devices consist of a vacuum tube containing a cathode for emitting electrons, an anode for attracting them, and (leaving diodes out of account) a grid for controlling their flow. But whereas valves are designed for using the electron stream (or anode current) externally, cathode-ray tubes use it inside by directing it against one end of the tube, where its effects can be seen. This controlled stream of electrons was originally known as a cathode ray; hence the name of the tube.

Fig. 23.1 shows a section of a c.r. tube. The collection of electrodes at the narrow end is termed the *gun*, because its purpose is to produce the stream of high-speed electrons. There is the cathode

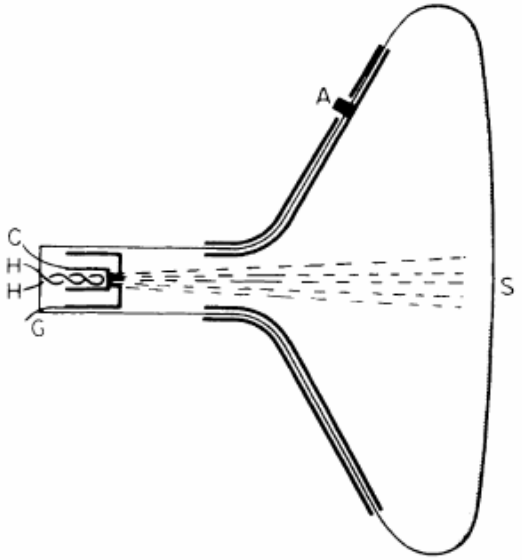


Fig. 23.1 Section of cathode-ray tube for television. H = heater; C = cathode; G = grid; A = anode; S = fluorescent screen

electrons from reaching the phosphor through it, and it is sufficiently mirror-like to reflect the glow and thereby increase its brightness as seen from outside the tube.

The only thing that can be done with the tube as described so far is to vary the brightness of the patch of light on the screen by varying the grid bias. To make it useful, two more things are needed: a means of *focusing* the beam of electrons so that the patch of light can be concentrated into a small spot; and a means of *deflecting* the beam so that the spot can be made to trace out any desired path or pattern. Both of these results are obtained through the influence of electric or magnetic fields. The fact that electrons are so infinitesimally light that they respond practically instantaneously, so that the trace of the spot on the screen faithfully portrays field variations corresponding to frequencies up to millions per second, is the reason for the great value of the cathode-ray tube, not only in television but in almost every branch of scientific and technical work.

In picture tubes both focusing and deflection can be done by magnetic fields, but as the electric methods are rather easier to understand we shall consider them first.

23.2 Electric Focusing and Deflection

When considering electric fields we visualised them by imagining lines of force mapping out the paths along which small charges such as electrons would begin to move under the influence of the field forces (Sec. 3.3). For example, if a difference of potential were maintained between two parallel circular plates, A and B in Fig. 23.2*a*, the lines would be somewhat as marked—parallel and

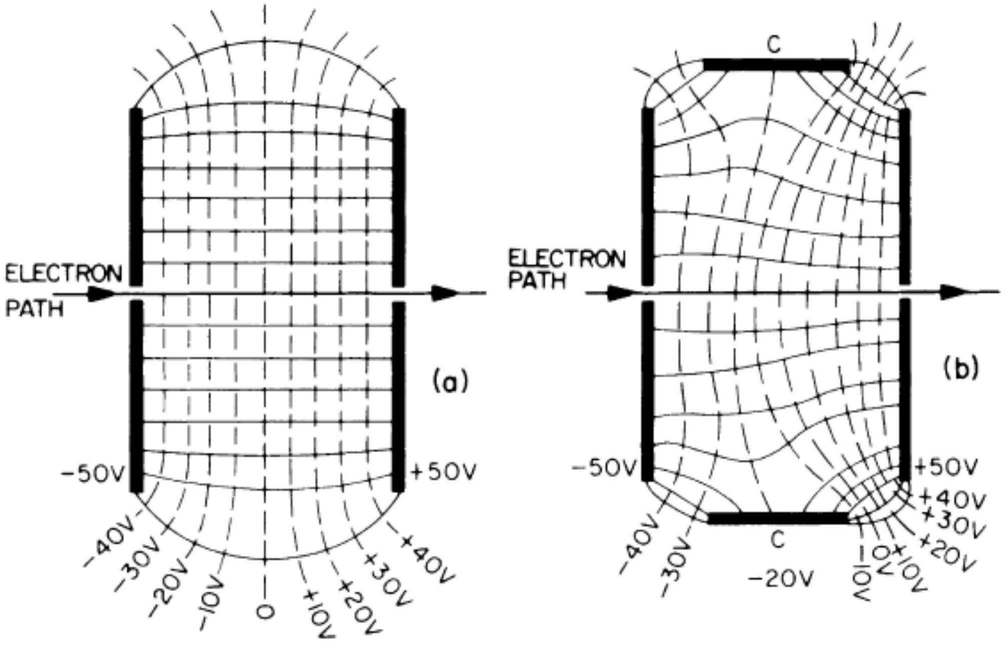


Fig. 23.2 Diagrams to illustrate electric focusing

really fine focus, so a good electron lens generally contains at least three anodes at different voltages. The focus is usually adjusted by varying one of the voltages. There are too many different varieties of electron lenses in c.r. tubes to illustrate even a selection, but Fig. 23.3 shows one of the simpler arrangements, with a potential divider for maintaining the electrodes at suitable voltages.

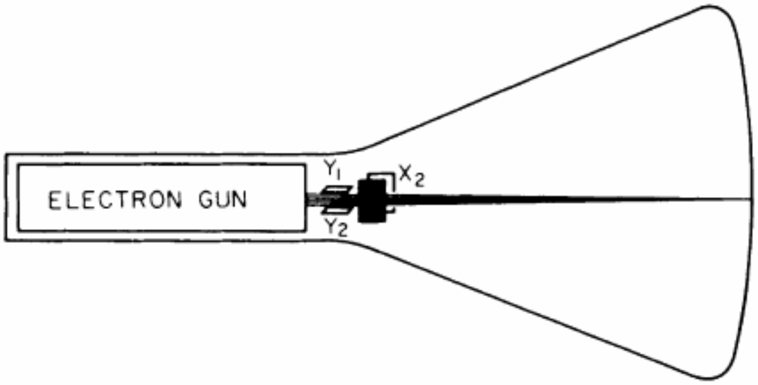


Fig. 23.4 Arrangement of deflector plates in a cathode-ray tube

Electric deflection is quite simple. If two parallel plates are placed, one above and the other below the beam as it emerges from the gun (Y_1 Y_2 in Fig. 23.4), and there is a difference of potential between the plates, the electric field between them will tend to make the electrons move from the negative to the positive plate. This tendency, combined with the original axial motion, will deflect the beam upwards or downwards, depending on which plate is positive. A second pair of plates, X_1X_2, is placed on each side of the beam for deflecting it sideways. By applying suitable voltages between Y_1 and Y_2 and between X_1 and X_2 the spot of light can be deflected to any point on the screen.

23.3 Magnetic Deflection and Focusing

Magnetic deflection depends on the force which acts on an electric current flowing through a magnetic field. In this case the current consists of the electron beam, and the field is produced by external coils close to the neck of the tube. The force acts at right angles to the directions of both the current and the field (Sec. 4.2), so if the coils are placed at the sides of the beam as in Fig. 23.5, so as to set up a magnetic field in the same direction as the electric field between the X plates in Fig. 23.4, the beam is deflected, not sideways, but up or down. A second pair of coils, above and below, provides sideways deflection.

With electric deflection, the angle through which the beam is deflected in inversely proportional to the final anode voltage. So if, say, 100 V between the plates is sufficient to deflect the spot to the edge of the screen when the voltage on the final anode is 1000, raising the anode voltage to 2000 makes it necessary to raise the deflecting voltage to 200. With magnetic deflection, the angle is inversely proportional to the square root of the anode voltage; so if, say, 20 mA in the deflecting coils was sufficient in the first case, it would have to be raised to only 28.3 mA in the second.

Raising the anode voltage thus necessitates more deflecting current or voltage, but it brightens the glow on the screen and generally improves the focus.

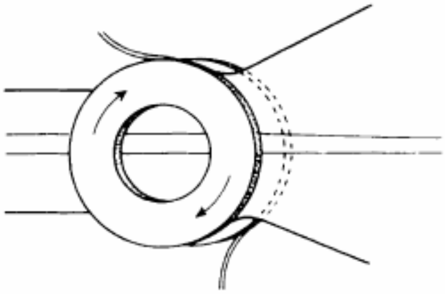


Fig. 23.5 Arrangement of deflector coils around a cathode-ray tube

For focusing magnetically, it is necessary for the direction of the field to be along the tube's axis. Electrons travelling exactly along the axis are therefore parallel to the field and experience no force. Those that stray from this narrow path find themselves cutting across the field and being deflected by it. That much is fairly obvious, but it is not at all easy to predict from first principles the final result of such deflection. Actually, by suitably adjusting the field strength, the electrons can be made to travel in spiral paths which converge to a focus in the desired plane. In picture tubes the magnetic field can be produced by a coil wound around the neck of the tube, in which case the focus is adjusted by varying the direct current through the coil. Alternatively one can use a ring-shaped permanent magnet, with a movable yoke for focus control. As with electric focusing, the beam is subjected to the focusing field before being deflected.

A slightly different principle is used in the long magnetic lenses employed with television camera tubes. The tube, including its deflecting coils, is surrounded for the whole of its length by the focusing coil. The electrons again pursue spiral paths which bring the beam to a focus at regular spacings along the length of the tube. By adjustment of the direct current in the focusing coil (or one of the accelerating voltages of the tube) one of the focus points can be made to coincide with the target. In long magnetic lenses focusing and deflection occur together.

23.4 Oscilloscopes

To the electronics engineer a cathode-ray oscilloscope is at least as television however deflection is practically always magnetic.

The usefulness of an oscilloscope would be very severely limited without amplifiers to enable adequate deflection to be obtained from weak signals. Ideally, such amplifiers would give a known constant linear gain at all frequencies from zero to the highest in any signals likely to be applied, and would introduce no phase shift. Such amplifiers have much in common with video amplifiers and their design, particularly where high gain is required, is not easy. In fact the principles of these amplifiers justify more space than can be given here and are therefore described in the next chapter. The cost of an oscilloscope depends in no small degree on the specifications of its amplifiers, and especially on their maximum frequency. On some of the cheaper models there is no X amplifier, the assumption being that either the time base will be used, or some other source of comparable voltage.

Because of the difficulties of providing high gain without appreciable distortion of any kind, high deflection sensitivity of the c.r. tube is desirable; but this conflicts with the essential requirements of good focus and bright spot trace, which call for high anode voltage and consequently low deflection sensitivity. Tubes for oscilloscopes therefore often include a feature called *post-deflection acceleration* (p.d.a.), a name that really explains itself. In such a tube the voltage on the anode that accelerates the electron beam ready to pass between the deflection plates may be as low as 1 kV, enabling good deflection sensitivity to be obtained; after deflection the beam is further accelerated by an anode held at 10 kV or even more. Sometimes the acceleration is imparted progressively by a very high-resistance spiral of carbon deposited on the inside of the flared portion of the tube and acting as a continuous potential divider between the final anode and the pre-deflection anode.

23.5 Time Bases

The ideal waveform for the time-base (X-plate) deflecting voltage (or current, if magnetic deflection is used) as described in the preceding section is shown in Fig. 23.6, which explains why it is called a sawtooth waveform. The most important feature is the

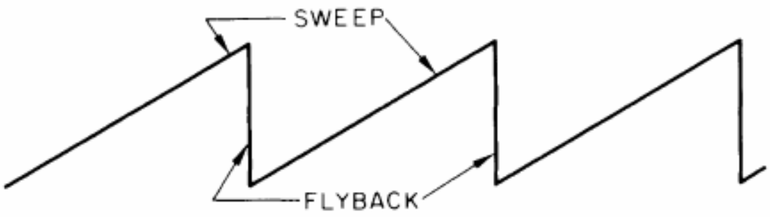


Fig. 23.6 Sawtooth waveform used for c.r. tube time-base deflection

To make the time base begin always at exactly the same phase in the signal cycle, so that a steady trace appears on the screen, some kind of synchronising device is needed. Here again, many techniques have been worked out. The simplest method is to arrange for a controllable amount of the signal voltage to be fed to the time-base generator at such a point that it prods it into action.

If magnetic deflection is used, as in television, a sawtooth current is required. The voltage waveform needed to cause such a current would itself have a sawtooth form only if the impedance of the deflection coils were pure resistance, but they are bound to have some inductance as well, and at high frequencies the inductance may predominate. As we saw in Fig. 4.10, the rate at which current grows in an inductance is proportional to the voltage across it. So if we want the current to grow at a constant rate we have to apply a constant voltage.

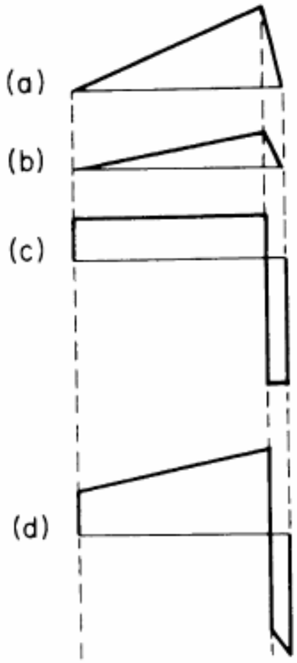


Fig. 23.8 (*a*) Required waveform of magnetic deflection coil current. The voltage waveform to overcome the coil resistance is shown at (*b*), the voltage for the coil inductance at (*c*), and the total voltage at (*d*)

In general the deflection coils have both resistance and inductance, so if *a* in Fig. 23.8 is the required deflection current the part of the applied voltage needed for the coil resistance would have the form shown at *b*, the part needed for the inductance at *c*, and the total applied voltage at *d*. Quite a large amount of ingenuity has been expended on means for economically generating acceptable approximations to the ideal waveforms (Sec. 23.10).

23.6 Application to Television

For television transmission, as we saw in Sec. 1.8, images are regarded as made up of small areas called picture elements. For high the vertical time base is 50 and flyback occurs after every 312½ lines. This subdivision of the scanning is called *interlacing*, and each scan making up half the complete picture is called a *field* (rather confusingly, because it has nothing to do with electric or magnetic fields).

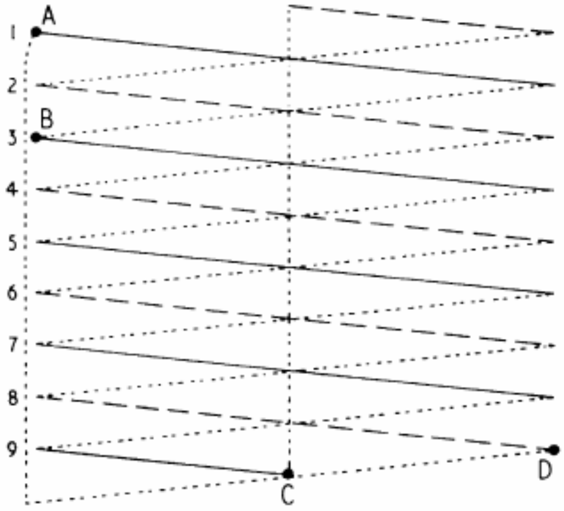


Fig. 23.9 Nine-line interlaced 'raster'. The dotted lines are flybacks, which normally are invisible

A scanning diagram showing all 625 lines would be difficult to follow, so Fig. 23.9 shows a *raster* (as it is called) of only 9 lines. The principle is the same, however. Beginning at A, the line time base scans line No. 1, at the end of which the flyback (shown dotted) brings the spot of light back to the left-hand edge. But not to A, for the field time base has meanwhile moved it downward two spaces to B. In the same way all the odd lines are scanned, until No. 9, half-way along which, at C, the field time base flies back to the top of the picture. Although for clearness it is shown as doing so vertically, this is not practicable as it would mean infinitely high speed. In a 625-line system a time of up to 31 lines is allowed, so fewer than 600 lines are actually available for the picture. During field and line flyback the cathode ray is cut off, so nothing appears. The line time base can now fill in all the even lines, shown dashed.

An essential addition to the system is some means for ensuring that all the receiving scanners work in exact synchronism with the scanner in the camera; otherwise the picture would be all muddled up. Synchronising signals are transmitted during the flyback periods of the camera scanner, and the receivers separate these from the picture signals and apply them to the time base generators to control their exact moments of flyback.

23.7 Characteristics of Television Signals

Two sample lines of combined signals, as used to modulate the

Either a.m. or f.m. can be used for television sound; in the 625-line standard it is f.m. The v.f. modulation is always a.m.

The specification shown faces a receiver with quite a difficult job,

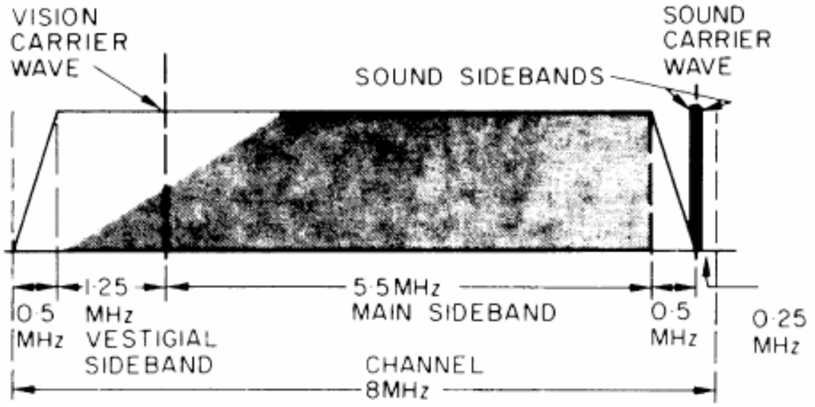


Fig. 23.13 Diagram of the frequency allocations in one complete 8 MHz television channel. The shaded portion is again the ideal receiver acceptance

for the vision section ought to respond as in Fig. 23.12 over a 6.75 MHz band and yet reject the associated sound only 0.5 MHz higher—and the adjacent-channel sound 0.75 MHz lower. Not infrequently receivers fail to do this, and the reproduced picture is visibly disturbed by the sound signal.

Ignoring the quite considerable problems at the sending end, let us now take a broad look at what a television receiver must include.

23.8 Television Receivers

Fig. 23.14 is a block diagram of a typical plain receiver for 625 lines. Usually there is a single preselector stage to ensure that the signal is strong enough when it comes to the relatively noisy frequency changer. We have already seen an example of this part of a television receiver; Fig. 22.16. It should be designed to amplify the full channel frequency band of 8 MHz equally; for example, from 470 to 478 MHz for the lowest-frequency u.h.f. channel, No. 21. If this is done, the adjacent-channel selectivity is likely to be hardly noticeable, but fortunately the range of reception on such frequencies is limited to the fairly local area and any other senders in that area would be allotted widely different frequencies. Change of channel at v.h.f. was usually obtained by switching different pre-tuned inductors into circuit, but with the transmission-line tuners used for u.h.f. this is not practicable and tuning is by ganged capacitor or varactor.

The frequency changer shifts the whole frequency pattern of the channel to the i.f., usually 33.5 MHz for vision carrier and 39.5 MHz for sound carrier. Because the amplifier has to cover such a wide band in relation to this i.f., the gain per stage is limited and several brightness. Moreover the polarity of the picture signals that is correct for the cathode normally happens also to be right for dealing with the synchronising signals. So cathode modulation of the beam is usual, the grid being reserved for control of general picture brightness by the viewer; there is no scarcity of voltage for that.

Because of the polarity of cathode modulation just referred to, the vision detector and v.f. amplifier must be arranged so that an increase in radio signal amplitude makes the cathode more positive, the 75% (black-level) amplitude being just enough to extinguish the spot altogether. The sync. pulses then make the cathode more positive still, so are unable to produce any visible effect, being 'blacker than black'. The same applies to a positive modulation system if the detector is connected the opposite way.

23.9 Synchronizing Circuits

A third outlet from the v.f. amplifier goes to the sync. separator, which removes the picture part of the signal, leaving only the sync. signals. It consists of what looks like a perfectly ordinary stage of resistance coupled amplification, Fig. 23.16 for example. As shown, the input is the video signal with the sync. pulses positive-going. (It would have to be reversed for a p-n-p transistor.) The d.c. component is removed by C, with the result that in a waveform diagram there will be equal areas above and below the zero-signal line, marked 0 in Fig. 23.15*a*, which shows two sample lines of video signal, the first corresponding to a bright scene and the second to a dark one. Thus the sync. pulses move up and down with scene brightness as shown in Fig. 23.15*a*, and a bias setting suitable for removing all but the sync. signals during a bright scene would remove the sync. signals too during a dark one. But it is possible to put this right by what is known as d.c. restorer action (Fig. 23.15*b*). R_1 in Fig. 23.16 is given an unusually high value, so that the base-to-emitter diode is on its bottom bend. It therefore rectifies the input signal as in a shunt-diode detector (Sec. 18.12), pushing it down until

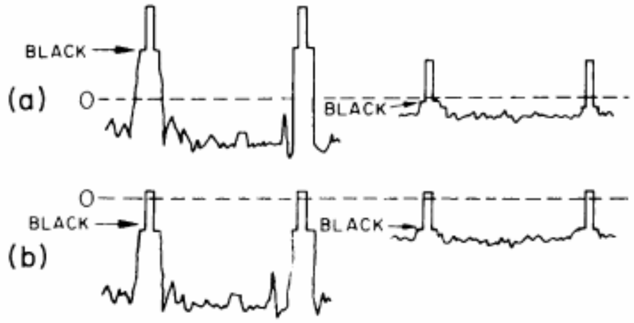


Fig. 23.15 Showing television signal voltage levels with (*a*) and without (*b*) d.c. restoration

into details one can say that the line time base is allowed to run continuously, producing pulses of its own at every flyback. These are compared with the incoming sync. signals in what is called a phase detector. If both occur at the same frequency, nothing happens. But directly the line generator gets out of step, the phase generator products a voltage, the polarity of which depends on whether the generator is running fast or slow. This voltage is applied to the generator in such a way as to correct its speed. The time constant of the system is chosen so that correction is gradual; occasional mutilation of the line sync. pulses by interference then has little or no effect, as there is an averaging out of the control.

23.10 Scanning Circuits

The line time base is the greater problem because of the greater amplitude required and its much higher frequency and the correspondingly short time—only a few microseconds—for the current through the deflection coils to be changed from maximum one way to maximum the other way during flyback. This is not remarkably fast by oscilloscope standards, but television is faced with the demand for maximum screen size combined with minimum distance from back to front. That means a very much larger angle through which the beam has to be deflected, and therefore a correspondingly larger effort from the time base generator (or scanner). Electric deflection is not practicable, so substantial *current* has to be reversed very rapidly. And as this is only one of many functions in a complete receiver, only a limited amount of power and space can be allowed for it.

The subject is extremely complex, and circuit diagrams of the many varieties of line scanner that have been used usually give little clue to the exact working conditions. In spite of the wide variations in detail, the basic principle is the same for all. Instead of providing a new lot of energy for each stroke—15 625 every second—energy is made to circulate to and fro, only the losses needing to be made up.

The basic method is so like the generation of continuous oscillation as described in Sec. 14.2 that it might be wise at this point to re-read that section. The example now to be described is the simplest of all, but it should be understood that its working has been idealised and that in practice there are many complications.

Fig. 23.18 shows the circuit. The type of transistor is one that can pass a heavy current, say 10 A, and withstand a relatively high collector voltage, say 150 V—but not both at the same time, as we shall see. Because it can pass current only in one direction, whereas the current in the deflection coils is required to alternate equally in both directions, a diode D is connected in parallel to provide a path for the opposite current. Although the line deflection coils are means for linearising, details of which can be found in books on television receiver design.

23.11 Colour Television

Reproduction of scenes in colour, by photography, printing and television, relies on the fact that any colour can be acceptably made up from three primary colours in the right proportions. The most apparently straightforward way of applying this fact to television is to transmit three complete pictures—one in each primary colour—in place of each one in the system we have just been studying. Assuming for the moment that pictures in the three separate primary colours (to each of which the term 'monochrome' *correctly* applies) can be reproduced on a single screen within 1/25th sec, 'persistence of vision' makes them appear collectively as a picture in full colour.

Although this scheme has been successfully demonstrated, the practical objection is that because everything has to be done three times as fast the bandwidth required is three times the already awkwardly wide requirement for plain television. Alternatively, the three primary colours could be broadcast concurrently on three channels, but the total frequency band would of course still be three times that for one ordinary channel. This factor of 3 could no doubt be cut down to something slightly less by taking advantage of the fact that with colour a rather lower standard of definition is acceptable. Even so, the economic disadvantages of this approach are so considerable that very much more complicated systems have been devised in order to avoid it and keep colour television within the standard channels. These are the systems in actual use. Although there are several, the differences between them are only in details which are too involved to include in a book of this kind. An essential requirement of any practical colour system is that it be *compatible*; that is to say, colour programmes must be receivable on non-colour receivers just as if they were uncoloured, and uncoloured programmes receivable as such on colour receivers.

The basic problem—to restate it in a different way—is to convey extra picture information (colour) without widening the frequency band. A solution might appear impossible, for a fundamental law of communication (Hartley's law) is that the product of bandwidth and time needed to convey a given amount of information is constant. So if you try to convey more information in a given time you have to use more bandwidth. This law assumes that the information is coded in the most economical manner. The compatibility requirement is the apparently insuperable difficulty, for it rules out any more economical coding of the basic picture (without colour), even if that were possible. So no room seems to be left for colour information.

The solution is to superimpose the colour information on the plain second gun strikes only green phosphor areas and the beam from the third gun strikes only blue phosphor areas. The use of a mask to screen each beam from the areas it must not strike is known as the shadowmask principle, and is used in all colour picture tubes although there are many different forms of construction. Fig. 23.24 shows one arrangement known as the precision-in-line (p.i.l.) tube, in which the phosphors are in the form of narrow vertical stripes on the screen (arranged in the order R, G, B), the shadowmask consists of vertical bars, and the electron guns are in a horizontal row, the beams passing through the apertures in the mask to strike the appropriate phosphor stripes.

23.12 Application to Radar

If we make a sound and hear an echo t seconds later, we know that the object reflecting the wave is 167 t metres distant, for sound travels through air at about 335 metres per second and in this case it does a double journey. Radar is based on the fact that radio waves also are reflected by such objects as aircraft, ships, buildings and coastlines. If the time taken for the echo to return to the point of origin is measured, the distance of the reflecting object is known to be 150 000t km away, the speed of radio waves through space being almost 300 000 km per second. Since distances which it is useful to measure may be less than a kilometre, it is obvious that one has to have means for measuring very small fractions of a second. In radar, as in television, time is measured in microseconds (μs).

This is where the cathode-ray tube again comes in useful. With a time base of even such a moderate frequency as 10 kHz one has a scale that can be read to less than 1 μs, and a much faster time base can easily be used if necessary. The time between sending out a wave and receiving its echo can be measured by making both events produce a deflection at right angles to the time base line. This can best be understood by considering a typical sequence of operations.

We begin with the spot at A in Fig. 23.25, just starting off on a stroke of the time base. Simultaneously a powerful sender is caused to radiate a burst or pulse of waves. This is picked up by a receiver on the same site, connected to the Y plates so that the spot is deflected, say to B. The pulse must be very short, so that echoes from the nearest objects to be detected do not arrive while it is still going on. In practice it may have to be a fraction of a microsecond, and as it has to consist of a reasonable number of r.f. cycles, their frequency obviously has to be many MHz. A more pressing reason for extremely high frequency is that it enables one to radiate the pulses precisely in any desired direction (Sec. 17.18). Typical frequencies are 3 GHz and 10 GHz (10 cm and 3 cm wavelengths).

The time base voltage continues to move the spot across the

CHAPTER 24

Non-sinusoidal Signal Amplification

24.1 Waveforms

Up to the last chapter, 'signals' were normally sinusoidal. That chapter was an introduction to applications in which sine waves hardly appear, except as radio carrier waves used to carry extremely non-sinusoidal forms. Most of these are of two main kinds. There is the pulse, which is often made to recur at regular intervals, as in Fig. 24.1*a*. If the pulses and intervals are equal, or nearly so, as at *b*, they are commonly called square waves, but in principle are the same. The next most important waveform is the ramp or triangle, which in recurrent form is commonly called sawtooth (*c*). We have already seen in Fig. 23.8*d* an example of how other waveforms can be compounded of these two. We have also had some inkling of the ease with which the sharp corners of such ideal forms as in Fig. 24.1 can be rubbed off in real circuits with their stray capacitances and inductances—even in circuits where sine waves are well preserved. On the other hand circuits suitable for pulses are quite likely to distort sine waves very seriously. So, having given so much attention to sine waves, we now need to look a little into the very different philosophy of sharp-cornered waveforms, the relative importance of which has for some time been increasing.

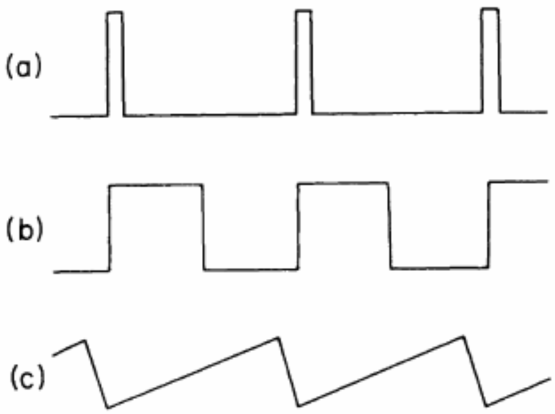


Fig. 24.1 Examples of the most important non-sinusoidal wave-forms: (*a*) narrow pulse, (*b*) square wave and (*c*) ramp or sawtooth

In spite of the very different principles and practice ruling sinusoidal and non-sinusoidal signals, there are connecting links in theory, and it is possible in practice to convert one to the other.

As far back as Fig. 1.2 we saw how, by adding together several greater. Triangular waves like Fig. 24.1*c* but with equal up and down slopes, on the contrary, are easier because the harmonic amplitudes decline inversely as the *square* of the harmonic number.

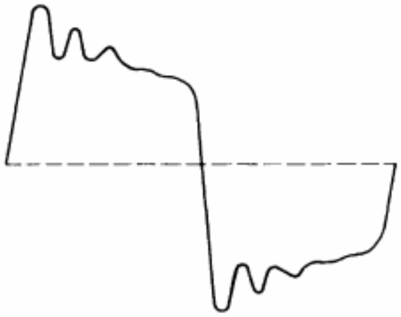


Fig. 24.3 Typical result of passing a nearly perfect square wave through circuit having restricted frequency coverage

The foregoing are just a few examples of the Fourier principle (Sec. 19.6) which states that *any* periodic (i.e., exactly repeating) waveform can be analysed into harmonic sine waves. Put in the reverse way, any periodic waveform can be synthesised by adding together sine waves, as we have been doing in Fig. 24.2 for the particular case of square waves. Reference books on signals often include formulae for the most useful waveforms. As with those we have been considering, they are all unending series, but one can stop at the point where a close enough approximation has been obtained.

24.2 Frequency and Time

By now it should have become clear that there are two alternative ways with signals. They can be treated either on a frequency or a time basis. If they are sinusoidal, frequency is much the more convenient. This particular waveform is unique in having only one frequency and in not being altered by any linear circuit composed of *R, C* and *L*. All the work we have done (notably in Chapters 6-8) on reactance, impedance, phasor diagrams, etc., may seem to have been quite hard—and it is only an introduction to much more advanced study of this kind—but it is dead simple compared with the mathematics of non-sinusoidal signals in circuits. That is one reason why the sine waveform is so much used. Reactance is based on frequency, and by enabling a.c. circuits to be calculated by an extended form of Ohm's law it provides a marvellous short cut. So much so that when dealing with non-sinusoidal signals it may actually be easiest (as we noted in Sec. 19.6) to analyse them into their component sine waves, find out what happens to these separately in the circuit, and then put the results together at the end. In general the output waveform will not be the same as at the input (Fig. 19.5). This procedure is like running round a tennis ball to avoid having to play it backhand. And in any case it relies on the circuit law (Super-position theorem) that the parts of a total current in a circuit can be calculated separately, not being affected by the other parts; and this is not true of non-linear communications (Hartley's law, Sec. 23.11). Now let us look at some of the basic problems encounterd in amplifying non-sinusoidal signals. The circuits to be described have applications other than the amplification of the waveforms illustrated in Fig. 24.1. For example they are used in oscilloscopes to amplify the input signal before application to the Y plates, and to amplify the time-base output to the level required by the X plates. Such amplifiers must have an excellent performance because the oscilloscope is used for precise measurements of the distortion in displayed waveforms. Another application for the circuits which follow is for amplifying video signals. The amplifiers used in radars, television studios and in video transmitting and recording equipment also need a very high standard of performance, but a lower standard can be accepted from the video amplifiers in television receivers because a small amount of certain types of distortion does not seriously degrade a television picture.

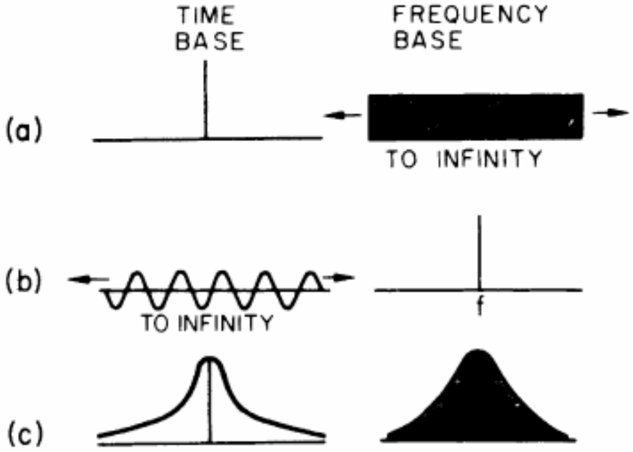


Fig. 24.5 Three different signal waveforms as they appear when graphed against time (left) and frequency (right)

24.3 Frequency Range

As already mentioned, if the edges of a pulse are truly vertical the number of harmonics is unlimited and the frequency range infinite. Clearly we cannot use (or even generate) such ideal pulses in practical equipment, which necessarily has a limited bandwidth. We are therefore interested in the effect on the shape of a pulse of eliminating all the harmonics above a particular frequency.

It is to tilt the edges so that they make a small angle to the vertical and to introduce curves at top and bottom as shown in Fig. 24.6. There is, in fact, a simple relationship between the steepness of the edges and the upper frequency limit. The steepness of the edges is measured by the time taken for the signal to rise from 10% to 90% of its final value. This is known as the rise time, and is related to the uppermost frequency f_{max} by the formula

$$\text{rise time} = \frac{1}{2f_{max}}.$$

If f_{max} = 10 MHz the rise time is 0.05 μs. There is a corresponding fall time, which is the time taken for the trailing edge to fall from 90% to 10% of the original value.

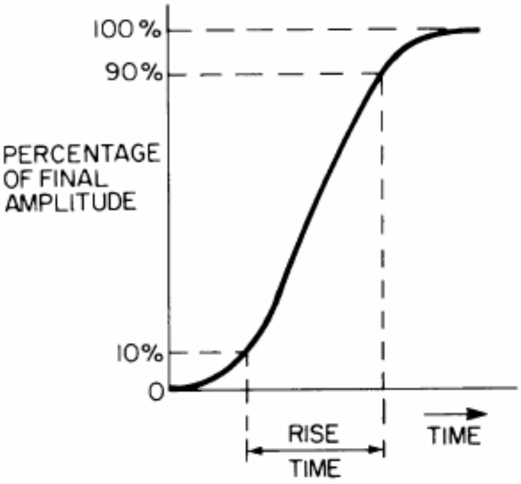


Fig. 24.6 The rise time of a pulse

As a good high-frequency response is needed to reproduce nearly vertical edges, perhaps one could guess that an extended low-frequency response is required to reproduce horizontal parts of waveforms. In fact, if a pulse switches between two truly horizontal levels, then the frequency response must extend down to zero frequency. In other words the amplifier must be capable of amplifying steady voltages. Such direct-coupled amplifiers are difficult to design as explained later (Sec. 24.6).

24.4 Importance of Phase Response

At a particular instant all the harmonics of a repetitive pulse signal are in phase so that the starts of their positive half-cycles all add to give the steep leading edge of the pulse. As these are being amplified they all suffer some phase shift as a result of the inevitable inductance and capacitance in parts and wiring. But at the output terminals the amplified harmonics must again be momentarily in phase so as to reconstitute the amplified steep edge. It is impossible to design an amplifier to have zero phase shift over a wide frequency band, and so instead we make the phase shift vary with frequency in such a way that all the harmonics have the phase relationship required to produce the near-vertical edge at the amplifier output.

To understand how this is achieved, consider the amplification of a single sinusoidal signal of frequency f and suppose that the frequency response and good phase response go together, and if the frequency response is made level over the band occupied by the pulse the phase response is likely to be satisfactory too. But the frequency response should fall off gradually above the upper frequency limit. A sharp cut-off above this limit will give an unsatisfactory phase response.

We have now examined the frequency and phase response of a pulse amplifier. A third important characteristic is the input/output relationship. For an a.f. amplifier this has to be as linear as we can make it, and negative feedback is extensively employed to achieve the required degree of linearity. This is also true of pulse and video amplifiers where very low distortion is essential. As already implied, such strict standards are not necessary in the video amplifiers of television receivers. For example if the frequency response of such an amplifier falls towards the upper cut-off frequency, the effect is to degrade fine detail in the reproduced picture, giving a soft effect described as poor definition. This can be compared with the effect of 'top cut' in an a.f. amplifier which robs the sound of its distinctness. Moreover strict proportionality between input and output signal may be undesirable. To appreciate this consider a device with a square-law characteristic, i.e., one in which the output is proportional to the square of the input. An a.f. signal which operated over the whole length of such a characteristic would produce very severe distortion at the output. But if a video signal is applied, all that happens is that signals near white level are amplified more than those near black level, i.e., contrast is increased. Provided no limiting occurs at either end of the characteristic, the detail in the scene is unaffected and many viewers might regard the output as an improvement on the input! Often the amplitude characteristic of a video amplifier is made deliberately non-linear in order to compensate for non-linearity in the characteristic of the picture tube, the light-output/signal-input curve of which is even steeper than a square law curve.

24.5 Obtaining the High-frequency Response

There is no difficulty in choosing transistors capable of uniform service up to 5 MHz; the main problem is the coupling impedance needed to develop the output voltage. We shall see in a moment that the greatest practical impedance is generally small compared with the output impedance of the transistor to which it is connected, so the gain can be regarded as proportional to the coupling impedance. It should therefore be constant over the whole working frequency band, and large enough for satisfactory gain.

Although transformers can be designed to cover a wide band of high frequencies *or* from say 30 Hz to 30 kHz, it is not practicable to

24.6 Operational Amplifiers

The shunt-injected feedback circuit (Fig. 19.22) was treated in Sec. 19.17 as a current amplifier. In practice a signal current source takes the form of a voltage source in series with such a high impedance that the current is determined almost entirely by it rather than by whatever circuit it works into. If this high impedance takes the form of a resistance that is a permanent part of the amplifier, the basis is laid for an interesting and important class of feedback voltage amplifiers.

Fig. 24.13 is essentially Fig. 19.22 plus this resistance, R_1. If the current gain is sufficiently large the part of the input current that goes into the amplifier at Y is negligible compared with what goes through R_2 to the 'live' output terminal, Z. For the feedback to be negative, the potential of Z has to be opposite to that of Y (relative to E). And because of the high gain of the amplifier it must be very much larger. The only way that these conditions can be fulfilled is for the signal potential to vary as shown in Fig. 24.14. If V_x alternates, the line see-saws about point P as the fulcrum, so the

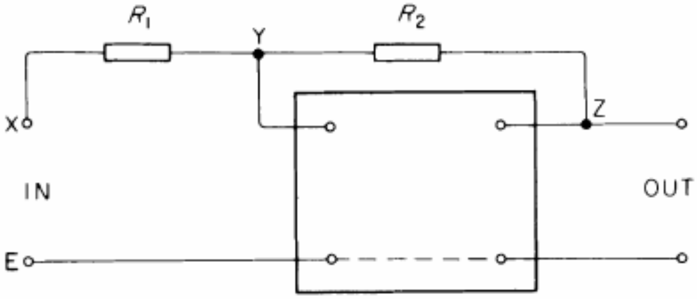


Fig. 24.13 The simplest form of operational amplifier circuit

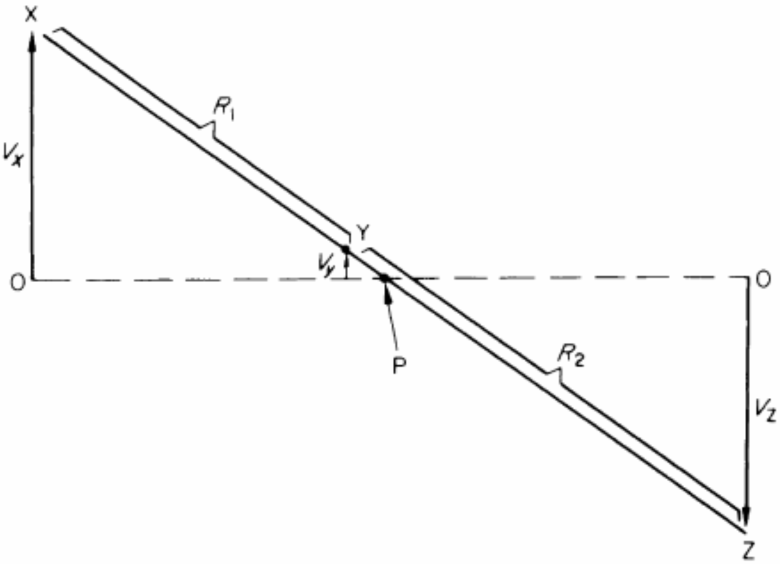


Fig. 24.14 Showing the distribution of potential along R_1R_2 in Fig. 24.13

potential as another to which it is not connected. (The virtual earth is a particular example.) In this way the effect of an unavoidable capacitance can sometimes be circumvented by preventing appreciable differences of potential from occurring across it.

24.7 Integrated Circuits

Because operational amplifiers consist of resistors and silicon transistors and diodes, and involve comparatively low power, the *integrated circuit* form of manufacture is ideal for them. A whole batch of identical amplifiers is laid down on a slice of silicon, the various p and n regions being produced by successive exposures to appropriate vapours which diffuse into the material. The conductive paths for resistors and 'wiring' are formed by metal deposits. The patterns of these actions are controlled by microscopic masks obtained by photographic reduction. Silicon dioxide prevents diffusion, so is laid down on the unmasked areas and can later be removed in preparation for the next process. The fact that pairs of transistors can be laid down very close together in identical circumstances minimises differences in performance that could result in common-mode effects.

A complete amplifier may occupy only one or two square millimetres. In fact miniaturisation has now reached the point where a single integrated circuit or 'silicon chip' (as the non-technical public call them) can contain as many as 100,000 components. The capital cost is large but the unit cost is small, so if large numbers are required they can be made cheaply. Inductors and large capacitors cannot be included, but reactances can be simulated by transistors with feedback phase-shifted by C and R (Sec. 19.21), and we have seen how negative feedback can be used to obtain precise performance from imprecise active devices. Design philosophy for integrated circuits is therefore quite different from that appropriate to circuits made up of separate components connected together with wire. But the principles on which they work are the same, so are covered by this book and there is no need to spend any more time on integrated circuits (i.cs) as such.

In practical applications i.cs usually have to take their places as components in systems where components of other kinds must also be included. If the minute size of the i.cs is not to be wasted, these other components should be as small as possible. An example of this is the ceramic filter described in Sec. 22.7. For compactness, stability and ease of production the wiring usually takes the form of a copper sheet laid down on insulating material (*substrate*) and eaten away by acid to the desired pattern (*printed circuit*) rather like the i.c. itself but on a large scale. On this substrate the components, including the i.cs, are mounted. Inductors of low value can be 'printed' on the sheet as spirals.

CHAPTER 25

Electronic Waveform Generators and Switches

25.1 Differentiating and Integrating Circuits

In the previous chapter we introduced the basic non-sinusoidal waveforms and we discussed the problems of amplifying them. In this chapter we continue the story of these waveforms by describing circuit techniques commonly used with them. We also consider methods of generating non-sinusoidal waveforms and introduce switching circuits—a most important type of circuit because they are used in vast numbers in, for example, digital watches, pocket calculators, television games and computers.

We begin by considering differentiating and integrating circuits. These names may seem rather formidable but we are already familiar with the things themselves. The names refer to mathematical operations. If y is some quantity that varies, differentiating is finding the *rate* at which it varies, often (but not necessarily) with time. Integrating is the opposite process: given the rate at which y varies, find y. In Fig. 25.1 on the top line we have a y that varies in a whole range of different ways for our instruction. Below it is plotted the rate at which y is varying; the mathematical shorthand for this, when the variation is with respect to time (t) is $\mathrm{d}y/\mathrm{d}t$. From a to b, y is increasing at the same steep rate all the time, so $\mathrm{d}y/\mathrm{d}t$ has a large positive value. From b to c, y is not changing at all, so $\mathrm{d}y/\mathrm{d}t$ is zero. From c to d it is decreasing steadily, at a slower rate than the increase from a to b, so $\mathrm{d}y/\mathrm{d}t$ is less and negative. de and ef are curves of the

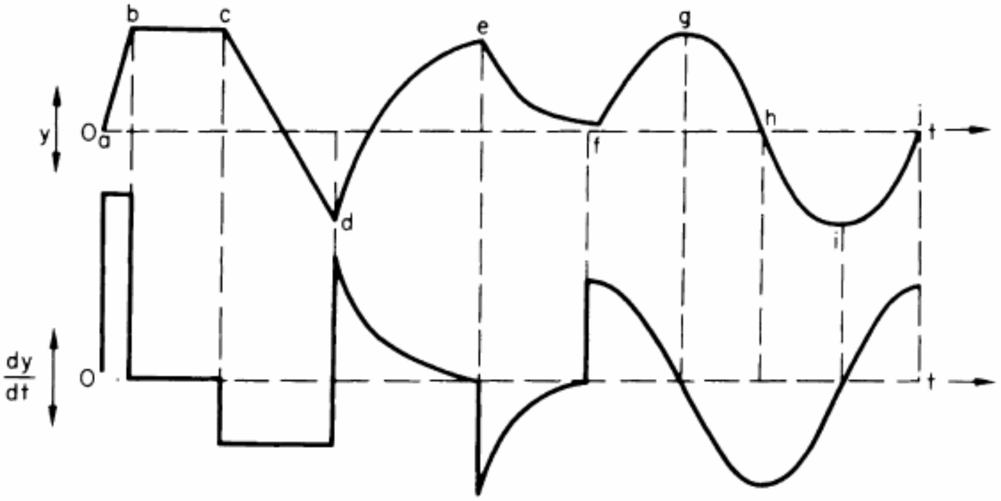


Fig. 25.1 The top line shows various ways in which a quantity can vary: below are the corresponding rates of variation

reversed, as in Fig. 25.5. This input, applied to *C* with its other terminal virtually earthed, ensures a constant input current. Flowing through *R*, this current sets up a constant output voltage, which of course is the differential of the input; and similarly any other input form is differentiated.

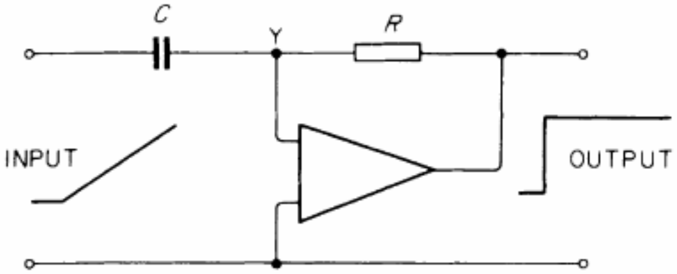


Fig. 25.5 And this one is the corresponding differentiating circuit

25.2 Waveform Shapers

The raw material for most non-sinusoidal waveforms—and sometimes sinusoidal ones—is the pulse or square waveform. It can be produced either by reshaping existing signals or by generating them direct by special oscillators. We consider shapers first.

Available signals are likely to be more or less sinusoidal—the mains supply, for example, or the product of an oscillator—and if the frequency is too low it can be multiplied (Sec. 19.5). They can easily be converted into nearly square waves by passing them through an amplifier that very readily overloads, or limits. Fig. 25.6 shows how simple this is. The positive half-cycles of the output voltage are all removed because the input voltage is then negative and allows only leakage current (negligible with silicon transistors)

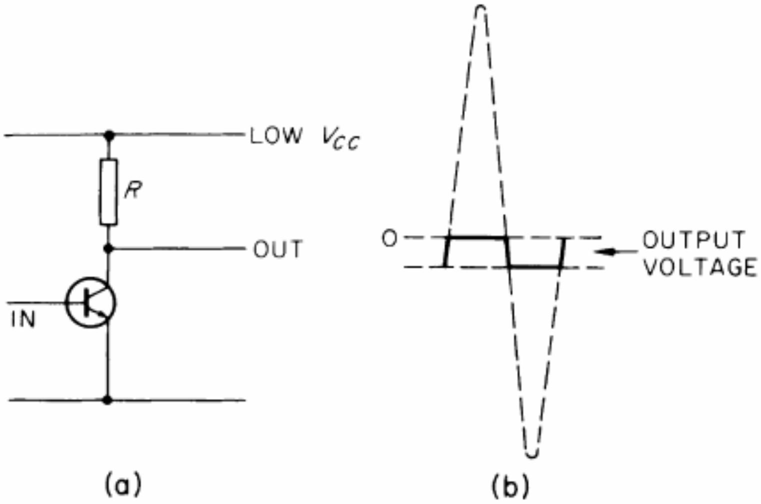


Fig. 25.6 Simple type of circuit (*a*) for obtaining a nearly square wave (*b*) from a sine wave

charged to nearly that voltage. When the pulse ceases, TR_1 is cut off and TR_2 is turned on, so C discharges through R_1 and TR_2. In the ordinary way the rate of discharge would fall off exponentially as the voltage of C fell, but owing to negative feedback Y is a virtual earth so the current through R_1—and therefore the discharge current, except only for some small variation in the base current of TR_2—is kept constant. To provide for a quick flyback, R_2 is much less than R_1.

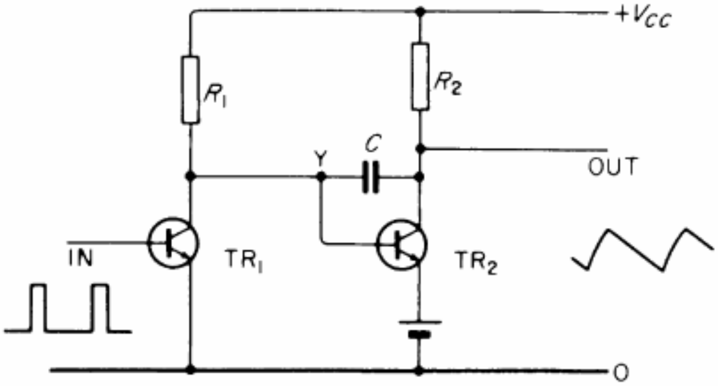


Fig. 25.9 A form of Blumlein integrator circuit, closely related to Fig. 25.4

25.3 Waveform Generators: Relaxation Oscillators

If a sinusoidal waveform is not required in addition to others developed from it as described in the last section, there is no point in carefully designing an oscillator for producing shapely sine waves only to distort them drastically. Oscillators can be designed to produce square waves or pulses directly; in fact, that is what an oscillator tends to do anyway if it is given its head. Besides dispensing with at least one stage, this method more readily yields a useful range of mark/space ratios.

When considering sine-wave oscillators we found that the most important conditions for least distortion of this waveform and for most precise frequency maintenance were the same: to use only just enough positive feedback to keep oscillation going. It follows that direct pulse generators, which depend on excessive positive feedback, are in general not so good at keeping themselves within very close frequency limits. That does not matter if, as in television, they can be synchronised from without. There is a whole series of types of pulse generator ranging from the completely independent to the completely dependent shaping types already discussed:

(1) Independent free-running oscillators establishing their own frequency.
(2) Oscillators which without control would operate in group 1, but which are pulled into step by external synchronising signals. 1 and 2 are both classed as *astable* systems.

(3) Oscillators which cannot run continuously on their own but which on being triggered from an external source will execute one complete cycle and then wait until the next triggering signal. This class is termed *monostable.*

(4) Oscillators similar to 3 except that when triggered they execute only part of a cycle and need another signal to make them complete it. These are called *bistable.* Although so dependent on outside sources, they do at least determine the duration and waveform of their part-cycles, and in this important respect differ from waveform shapers.

All these groups of oscillators are distinguished from the sinewave variety, in which the transistor works continuously as an amplifier, as *relaxation* oscillators, in which it works intermittently as a quick-acting switch. The frequency is determined not by an LC tuned circuit but by delays due to time constants, either CR or L/C (Secs. 3.5 and 4.8) or both. It is the relaxation or gradual discharge of the time-constant elements between the active or switching phases that gave rise to the name for this class of circuits. For an understanding we must once again keep in mind the fundamental principle that the voltage across a capacitor cannot change abruptly but only as a result of current flowing into or out of it for a time, and its dual principle, that the current in an inductor cannot change abruptly but only as a result of voltage across it for a time. Although all relaxation oscillators are basically the same in principle, having a large amount of positive feedback coupling through a time-delay circuit, one particular variety—the *blocking oscillator*—is usually distinguished from the rest.

25.4 The Blocking Oscillator

The distinguishing feature is inductive coupling, which makes blocking oscillators look deceptively like the ordinary oscillators described in Chapter 14; see for example Fig. 25.10. Attempts, in the interests of brevity, to explain their working along the same lines as ordinary oscillators are likely to mislead, for it is really quite different.

The cycle of oscillation comprises two phases, one of them being a small fraction of the whole period, in which a lot of things are happening very quickly, so let us begin with the other by supposing that C is charged so as to make point a in Fig. 25.10*a* negative. As this point is connected through R to $+V_{CC}$, C gradually discharges as shown by the exponential curve of v_a, which would eventually carry it to the V_{CC} level. The transistor being cut off by this negative bias, the collector potential v_c is the same as V_{CC}, while V_b is the same as v_a. Directly v_a becomes slightly positive, collector current begins to flow through the primary coil of the transformer. The rate of followed by a relatively long interval determined by CR, during which v_a and v_b are the first half of an exponential curve. The p.r.f. can therefore be conveniently controlled by varying R.

Because R tends to bias the transistor on, this form of blocking oscillator is free-running. But if a synchronising signal is injected at point a, just before the pulse is due, v_b crosses the zero line at once and triggers the transistor into pulse action. To ensure synchronism the free-running p.r.f. is adjusted to be slightly lower than the sync. pulse frequency.

Note that oscillators such as this, whose frequency depends on a time constant, can be synchronised not only to the frequency of the sync. pulses but to submultiple frequencies. Fig. 25.11 shows the graph of v_b again, with regular sync. pulses superimposed. It is not until the fifth of these that the voltage is able to rise above zero and start the pulse phase. The oscillator is here acting as a frequency divider.

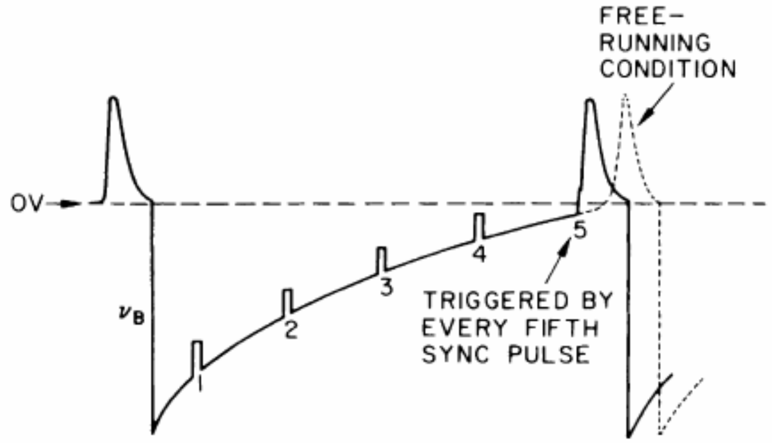


Fig. 25.11 Showing combined synchronizing and frequency-dividing action

This circuit or variations of it are commonly used in television receivers to generate the field scanning waveform. It must be followed by an output stage, with provision for linearising the sawtooth.

The circuit shown can easily be converted into a monostable one by connecting R across C, or in some other way arranging that the normal base-emitter voltage is zero, so that no action takes place until a triggering pulse is received, of the polarity that turns the transistor on.

25.5 The Multivibrator

The blocking oscillator is a remarkably simple device for obtaining very brief pulses. Transformers do not however fit well into modern circuit construction (Sec. 24.7) so for square waves or longer pulses

Because TR_1 is bottomed, v_{c1} is only just above zero. And because C_1 has had no time to charge, v_{b2} is pushed down to the same extent. Although in doing so it cuts off TR_2, v_{c2} cannot instantly spring up to $+V_{CC}$, because C_2 has to charge through R_2 and TR_1 (base to emitter). So v_{c2} rises exponentially as shown. v_{b1} is thereby pushed positive, but only slightly and temporarily because of the charging of C_2. When that is complete v_{b1} is held at a small positive bias by current via R_4. This fall-off in v_{b1} reduces the TR_1 collector current slightly, so there is a slight rise in v_{c1}. Meanwhile C_1, held on the left at the near-zero v_{c1}, is charging through R_3, so v_{b2} rises exponentially. Assuming, as is usual, R_3 and R_4 are considerably greater than R_1 and R_2, this growth is comparatively slow.

The end of Phase 1 is brought about by v_{b2} (similarly to v_b in Fig. 25.10) becoming positive, starting current in TR_2 and thereby a negative-going v_{c2}, which is communicated by C_2 to the base of TR_1, making v_{c1} rise. The amplification ensures a rapid reversal of conditions, TR_1 being switched off and TR_2 on. By the end of Phase 2 a complete cycle has been accomplished and the whole process repeats continuously.

Phases 1 and 2 being equal indicates that the component values are the same for each half of the circuit, or at least that $C_1R_3 = C_2R_4$, these time constants clearly being mainly responsible for the durations of the phases. Any desired mark/space ratio within reason can be obtained by choosing component values to make these time constants suitably unequal.

Like the blocking oscillator, the multivibrator is naturally free-running but can easily be synchronised by pulses injected almost anywhere in the circuit—usually at one of the bases. Both types of circuit are also self-generators of non-linear ramps. By making R_1 and R_2 as low as possible, a nearly square waveform can be obtained from either collector, and if necessary it can easily be improved by shapers (Sec. 25.2).

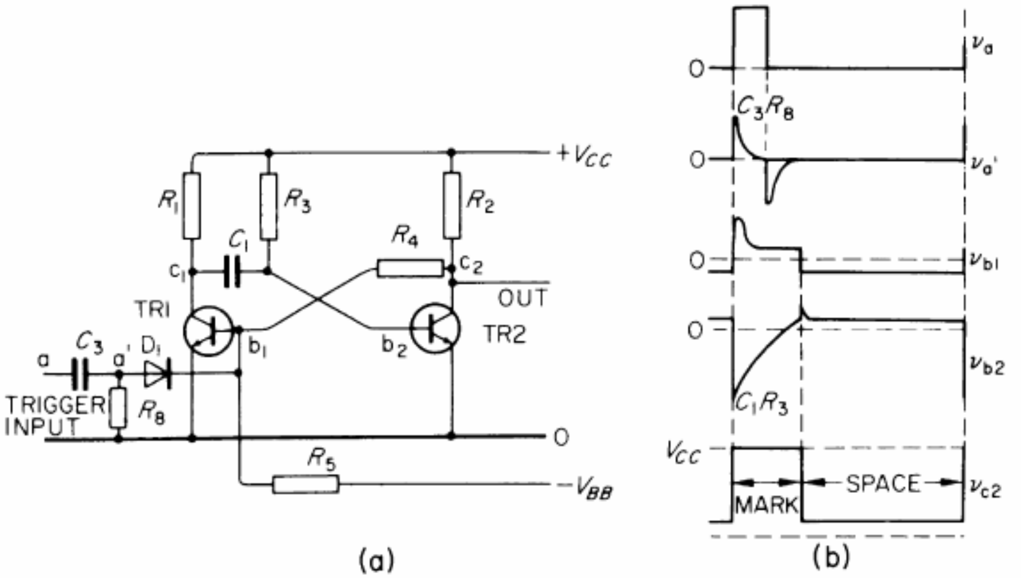
25.6 Bistables

If the two capacitors in the astable multivibrator are replaced by the base-biasing resistors as in Fig. 25.13 the circuit ceases to have a self-triggering action and has to be externally triggered both ways. In other words it is stable in both conditions (either transistor conducting) so is termed bistable. It has ceased altogether to be a self-oscillator, and in its basic form has no time constants; these, we recall, served as delay-action self-triggers.

Like the astable, this particular type of bistable was invented as far back as 1918, in valve form of course. It was named after the inventors as the Eccles-Jordan relay, and is now one of the main 'building blocks' in digital computers. Whereas the multivibrator is

25.7 Monostables

The monostable multivibrator comes half-way between the astable multivibrator, which can trigger itself both ways continuously, and the bistable, which can trigger itself neither way so stays put until it is triggered externally. It needs to be triggered from its stable condition, but triggers itself back to the original state after an interval determined by a built-in time constant. So we are not surprised to find that one half of the basic circuit is the same as in the astable type and the other half is the same as in the bistable type. Fig. 25.16 shows quite a practical form of the circuit, in which the negative bias to ensure that TR_1 remains completely cut off when in the stable state is supplied from a negative source instead of by the voltage drop due to emitter current through R_7 in Figs. 25.14 and 15. A suitable triggering circuit is also shown. For easy comparison with the previous circuits the corresponding components have the same numbers.



(a) (b)

Fig. 25.16 Practical form of monostable circuit with trigger (*a*), and its waveforms (*b*)

Typical waveforms are shown. A sample triggering pulse appears as v_a; its shape is not very important so long as the rising front is reasonably steep. It is differentiated by C_3R_8 which has a relatively short time constant to give the short sharp pulses v_a'. Only the positive one (v_{b1}) reaches the base of TR_1, the negative one being blocked by the diode D_1. The switch-over started by such a positive pulse sends b_2 sharply negative (see v_{b2}) and correspondingly c_2 sharply positive to form the front of the output waveform. Note that because there is no capacitor C_2 this front rises steeply instead of in a charging curve as in Fig. 25.10. The positive voltage thus imparted to the top end of R_4 maintains v_{b1} positive and TR_1 on after the triggering pulse has died away. As soon as C_1 has charged sufficiently to make v_{b2} positive, the switch-over from TR_1 to TR_2 takes place in the same manner as in the astable multivibrator.

There are many uses for this kind of circuit, such as generating pulses of controllable length, reshaping pulses that have deteriorated in transmission, and creating adjustable time delays.

25.8 Clampers and Clippers

Fig. 25.17 shows a simple multi-purpose device that we have already met more than once. It goes by various names according to the purpose. One of these, known as clamping, is shifting a wave-train as a whole, lining up either positive or negative peaks to any desired potential. As it is shown it lines up or clamps the positive peaks to what is arbitrarily regarded as zero voltage or earth (0 V). To clamp the negative peaks, reverse the diode. To clamp to some other potential, bias the lower end of the diode to that voltage. The action is precisely as already described in Sec. 18.12. The upward slope of the negative peaks of the output is due to the negative charge on C, accumulated through the diode conducting at the positive peaks, leaking off slightly through R. To minimise distortion of the waveform due to such leakage, the time constant CR should be many times greater than the period of one cycle. While the scheme works more or less with any waveform, one that is very sharply peaked at the polarity of clamping (such as the sync. pulses being treated in this way in Fig. 23.15) is likely to have its peaks clipped somewhat because of the very short time available for topping up C.

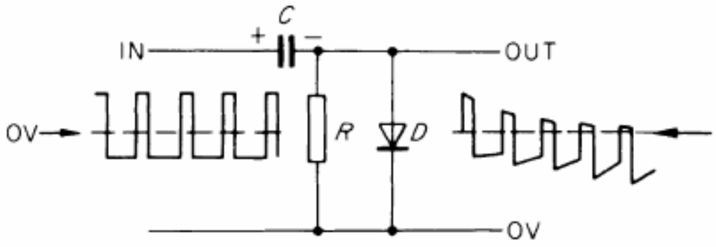


Fig. 25.17 Diode clamping circuit

If the wave-train is amplitude-modulated, the time constant CR has to be more carefully chosen, according to what is required. If it is made sufficiently long the charge will remain in C not just between one cycle and the next, but between one maximum-amplitude cycle and the next, with perhaps many weaker cycles in between. Fig. 25.18 shows this, with modulated sine waves clamped positive of zero. CR in this case must be long compared with T, which is the length of one cycle of the *modulation* frequency.

Note that this output contains the original modulated carrier wave plus a z.f. (d.c.) component. So, as with the unmodulated pulses, the represent suitable sources of bias. Until the instantaneous signal voltage exceeds + V_1 or V_2 neither diode is conducting and the arrangement is a potential divider reducing the signal in the ratio

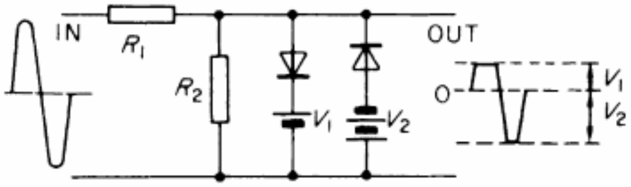
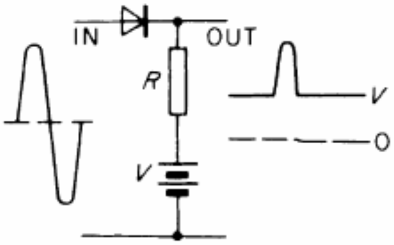


Fig. 25.20 Example of clipping circuit for both polarities

$R_2/R_1 + R_2$). R_2 is not strictly necessary as a component but represents the resistance of whatever is being fed, combined with the leakage of the diodes if that is appreciable. R_1 should therefore not be larger than it need be to make the resistance of the diodes when conducting, plus their bias sources, negligible in comparison with it, so that an appreciable fraction of what is supposed to be removed is not developed across them.

Fig. 25.21 Series-diode type of clipping circuit



That is the shunt type of clipper. The series type (Fig. 25.21) is more suitable for performing the reverse operation: rejecting all of a waveform that does not exceed the bias voltage. Note that this voltage appears at the output all the time. With no bias this clipper is a simple rectifier giving a mean output equal to half the mean value of a half-cycle of the input (Sec. 27.3).

25.9 Gates and Choppers

In computers and other electronic equipment much use is made of electronic switches or *gates*. They fall into two main classes: *linear* or transmission gates, in which the signal gets through in its original form when the gate is open and not at all when it is closed; and *logic* gates, in which the output is due to the combination of signals present at the input and is not necessarily of the same form.

Diodes can be used as linear gates, but as many as six may be needed in specially balanced circuits in order to keep the controlling and controlled signals separate without serious loss of signal. This problem is reduced by using transistors, in which the base is available as a control electrode to render, the collector-emitter path

Transistors can also be used in OR and AND gates with the advantage of gain as well if that is needed. Another facility provided by transistors and much used in logic circuits is inversion, which changes 'signal' to 'no signal' and vice versa. Although any common-emitter amplifier inverts, so hardly qualifies as a gate, in the context of electronic logic it is called a NOT gate. Most often however it combines the NOT function with being an OR or AND gate, the combination being termed a NOR or NAND gate.

A simple n-p-n-transistor gate is shown in Fig. 25.24 with the relevant waveforms. With both A and B at zero volts the base is sufficiently negative to cut off the transistor and give + 12V output. If either A or B input is made not less than about + 3V, the base goes slightly positive, making the transistor conduct and bring the output voltage near zero. To reduce interaction between the input sources and also to reduce loading on them, the input resistors may be replaced by diodes as in Fig. 25.23*a*. For the positive logic convention this gate gives an inverted or NOT signal whenever there is an input at A or B; this is the property of a NOR gate. For negative logic, however, the gate gives a NOT signal only when there is an input at A and B—the property of a NAND gate. So, as expected, the behaviour of this gate depends on the logic convention.

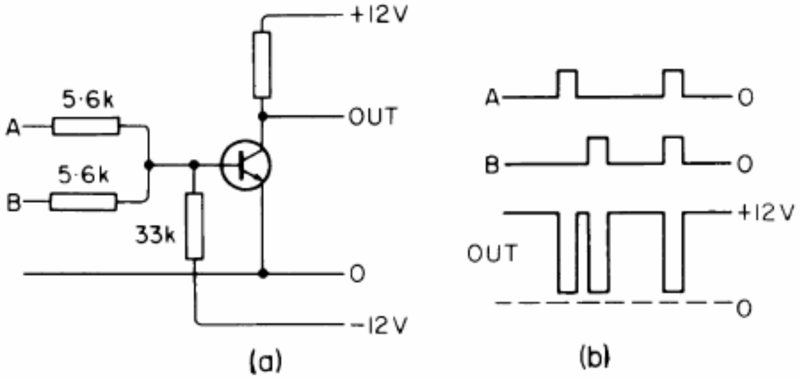


Fig. 25.24 A transistor NOR gate; a + 12V output is obtained whenever there is a pulse at neither A nor B

If a p-n-p transistor is substituted in Fig. 25.24 and all the polarities consequently reversed, the result is an AND gate for positive logic and a NOR gate for negative logic. If any input is at zero volts the base is slightly negative which makes a p-n-p transistor conduct and keeps the output voltage only slightly less positive than 0V. Only if all inputs are positive can the base go positive and cut off the transistor, making the output voltage fully negative.

25.10 Design

You may be disappointed not to find component values given for all the circuits in this chapter, or at least detailed instructions on how to

CHAPTER 26

Computers

26.1 Analogue and Digital

It would of course be absurd to attempt to cover the subject of computers in one short chapter. All that can be done is to indicate broadly what they involve and how they make use of the principles we have been studying.

The first thing is to distinguish between *analogue* and *digital* computers. In analogue computers the magnitudes being processed are represented by continuously variable quantities, such as voltage, current or charge, whereas in digital computers they are represented by numbers, which are not continuously variable. One third cannot be expressed by any finite number of decimal digits (being 0.333 . . . . for ever) but one third of a distance is theoretically possible. Paradoxically, however, digital computers and calculators are capable in practice of far higher precision than analogue types.

Electrical instruments such as voltmeters illustrate the distinction. The ordinary pointer types are analogue instruments, voltage being indicated in terms of a different quantity: the *angle* of the pointer. Such voltmeters are simple in principle and have comparatively few parts, but the standard of workmanship goes up so steeply with the degree of precision demanded that they are limited in practice to very moderate precision. Digital voltmeters, in which the reading is exhibited in figures, are very much more complicated and contain very many parts, so for ordinary 1% accuracy cannot compete in price with pointer instruments. But the number of kinds of parts is quite small and they are individually cheap, and the precision of reading can be increased several decimal places by increasing the number of parts of the same few kinds, little or no improvement in quality being necessary.

The automatic telephone is a digital system. Seven twists of its simple mass-produced dial, generating seven groups of up to ten pulses, could be used to select any one from ten million different numbers. An analogue scale capable of being read unmistakably to one in ten million would be a formidable task for the instrument maker!

Coming to actual computers, we may consider that one-time very familiar analogue type, the slide rule, in which numbers are represented by lengths along scales, proportional to the logarithms of the numbers. A digital calculator to perform to the same standard, which is sufficient for most engineering purposes, would be somewhat more expensive. But it could work incomparably faster. And digital calculators can be made incomparably more precise.

26.2 Analogue Computers

Electronic analogue computers are particularly useful for engineering purposes, for several reasons: few engineering data are known, or need be known, very precisely (think, for example of the inevitable tolerances in transistors, and the uncertainty of maximum loading on a structure); quantities such as temperature, pressure and velocity can easily be 'transduced' into proportional voltages or currents instead of first into numbers; and the effects of varying the data can be observed continuously. The engineering designer used to have to arrive at his specifications by laboriously calculating for several alternative values of each parameter and choosing the best combination. Now it is possible to vary them and see the results at once, notwithstanding that the mathematical relationships may be quite complex. Or he may even hand the whole problem over to the computer and get it to indicate the optimum specification.

The same techniques are involved in the many kinds of simulators, such as those used for the early stages of learning to fly or to drive, without the expense and risk of using actual aircraft in flight or cars on the road with novices in control.

The basic components of analogue computers are operational amplifiers, to which we were introduced in Sec 24.6. They are available as microscopic integrated circuits (Sec. 24.7) so although large quantities of them are needed they occupy little space. They use voltage as analogues of whatever numerical data are being computed. An inverter amplifier (Fig. 24.16*a*) changes the mathematical sign; a potential divider (up or down) multiplies or divides. If several input voltages are applied, each through its own resistance equal to R_1 in Fig. 24.16*a*), the total input current at Y is equal to the sum of the separate input currents, so the output voltage is equal to minus the sum of the applied voltages. That gives addition—and subtraction, as negative voltages can be applied. Factors can be introduced by using different values of R_1. Differentiation is accomplished by substituting a capacitor for R_1 and integration by substituting a capacitor for R_2. (Sec. 25.1). By suitable combinations of operational amplifiers, differential equations are almost instantly solved. Waveform generators or shapers are sometimes used to introduce mathematical functions corresponding to them.

The basic disadvantage of analogue computers is that any departure from exact proportionality of the voltages to the quantities represented by them introduces error. It is only the use of large amounts of negative feedback which, by nearly eliminating the effects of variations in the active devices in the operational amplifiers, keeps such errors within reasonable limits.

26.3 Digital Computers

In order to understand what a digital computer has inside it and why, let us consider what we do when we compute mentally or with the aid of pencil and paper. First we have to receive the data and instructions on what to do with them; for example, 'Add 27 to 189 and multiply the result by 17'. The first part of a computer, correspondingly, is an *input unit*, to receive data and instructions in a form the computer can use, such as punched cards, punched or magnetic tape, teleprinter signals, typewriter keys, etc.

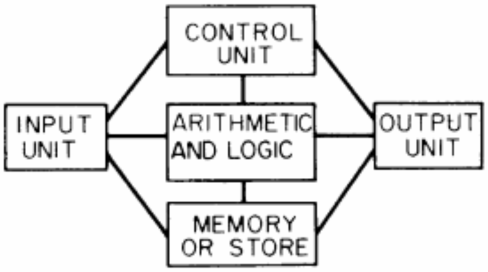


Fig. 26.1 A computer can be divided into these five departments

Next, we perform the arithmetical operations of adding and multiplying. The regular procedure with the above problem would be to add 7 to 9, get 16, put down 6 in the units column, carry 1 to the tens column and add it to 2 and 8, getting 11, put down 1 and carry 1 to the hundreds column, adding it to 1 to get 2. Partial answer: 216. Then we would multiply this 216 by 7 and add it to 2160. Or we might have a × 17 multiplication table or ready reckoner giving the answer directly. The part of a computer that manipulates the given digits to provide a partial or final answer is called the *arithmetic and logic unit*. The part that uses the given instructions to bring this unit into action in the appropriate ways at the right moments is the *control unit*, and the part that stores partial results until they are needed (taking the place of the paper on which we write them down, or our memory if it is good enough) and such useful information as the multiplication table to shorten frequently-needed arithmetical operations, is called the *store* or *memory*.

Finally, the part that converts the answers from computer form to a form that is convenient for the user of the computer, is the *output unit*. Its 'readout' may be displayed as printed numbers or words, or signals to control a machine.

These divisions are shown diagrammatically in Fig. 26.1.

If the computer is a general-purpose one of moderate capability, with plenty in the arithmetic and logic unit but not much in the control unit, the instructions (called the *program*) have to be given in considerable detail, so the services of an expert programmer are needed. Very elaborate computers, or special-purpose ones, have programs built into them. They work rather like a military organisation in which it is not necessary to give a complex multitude of orders in detail to bring about, say, an invasion, but only a simple 'Operate Plan B' which has already been worked out. These programs are stored in the memory.

A number of different types of store are used, sometimes more than one type in a single computer, depending on whether the stored information is much used and has to be found very quickly, or not. Most stores are magnetic, using tapes, drums or disks, or are electronic in the form of a matrix of transistors in an i.c.

26.4 The Binary Scale

We normally use a decimal system of numbers, presumably because we normally have 10 fingers (including thumbs) on which to count. A primitive tribe is reputed to use only three numbers: one; two; plenty. When the need arose for numbers greater than a reasonable variety of symbols, the simple and ingenious convention was adopted of placing a symbol one space to the left to mean the number of groups of ten, and a further space to the left to mean the number of ten groups of ten (i.e., hundreds); and so on. Using only ten different symbols, this system extends the range of numbers indefinitely. We may be so used to it that we hardly realise that 47023 (to take an example of a decimal number) is really shorthand for

$$(3 \times 10^{0}) + (2 \times 10^{1}) + (0 \times 10^{2}) + (7 \times 10^{3}) + (4 \times 10^{4})$$

And of course we shift to the right for subdivisions, 10^{-1}, 10^{-2}, etc. the units position ($\times\ 10^{\circ}$ or 1) being identified by a dot on its right-hand side.

For the purposes of digital computers this system has the disadvantage of necessitating ten different unmistakable electronic states, such as values of collector current, to represent the ten numerals. Even if extensive negative feedback is used to counteract changes in transistor characteristics and supply voltages, it is very difficult to alter the states at will by the required number of units without risk of error. And even one unit of error is serious when it is in the millions column!

But if the number of different states is reduced from ten to two everything becomes very easy. We need only consider two states: 'off' and 'on'. The last chapter was mainly about circuits that are in an 'off' or an 'on' state, and signals that can be used to change can however be made for the computer to do it directly. 'Long multiplication' is extremely simple in the binary scale. For example, multiply 11010 by 101:

$$\begin{array}{r}
11010\\
00000\\
11010\\
\hline
10000010
\end{array}$$

In practice one would not write down the second row, or any row corresponding to 0 in the multiplier. As we see, the essential processes are shifting the whole multiplicand successively to the left and adding all the results. Means for doing both these things will appear in the next section.

There are several possible electronic methods of subtracting, corresponding to methods people use. In one, we replace each digit of the number to be subtracted by its nines complement—the number needed to make it up to nine—and *add*, finally transferring the 1 at the left-hand end of the right-hand end; for example, take 108 from 762:

$$\begin{array}{r l}
762 & \\
891 & \text{(nines complement of 108)}\\
\hline
1653 & \text{(sum)}\\
\llcorner\!\!\_\!\_\!\_\!\!\uparrow & \text{Answer: 654}
\end{array}$$

A binary computer similarly adds the ones complement, which it obtains simply by means of a NOT gate (Sec. 25.9); for example, take 1101 from 10110:

$$\begin{array}{r l}
10110 & \\
0010 & \text{(ones complement of 1101)}\\
\hline
11000 & \text{(sum)}\\
\llcorner\!\!\_\!\_\!\_\!\_\!\!\uparrow & \text{Answer: 1001}
\end{array}$$

Division is repeated subtraction. To take a decimal example, 34 ÷ 9; 9 can be taken from 34 three times, leaving 7 over, so the answer is $3\frac{7}{9}$. The binary computer has to count the number of times the subtraction can be done before the remainder becomes negative.

26.5 Logic Circuits

Obviously we need a set of symbols for the logic gates described in Sec. 25.9, not only for brevity but also for clarity because the circuit with what you get by doing the sum mentally. They should of course be the same.

Another essential component of any arithmetic and logic unit is the shift register. This is a row of as many bistables as there are bits in a word. Its purpose is to transfer the bits one place to the right every time a regular periodical shift pulse occurs. In this way a whole word can be stored. Fig. 26.5 shows this function for a simple example of only five digit places. The number, 10011, is first shown ready to enter, at the left. One pulse from the clock (Sec. 26.8) transfers the units digit to the first bistable, where it exists as one of its two possible states. Which one it is determines what happens as a result of the next shift pulse. If it is an 'on' or 1 state, as here, it reverses the next bistable from 0 to 1. And so in successive stages the whole word is taken in. The last diagram shows the first digit emerging as a 1 pulse. If the last bistable's output is fed back to the first, as in some shift registers, the word is kept circulating and is thus stored until required and made to emerge by some control signal. Or it may be directly applied to the *A* input of an adder and added one digit at a time to another numerical word now coming in directly, or shifted one place at a time for multiplication.

For its various functions the basic bistable circuit is subject to slight modifications. The one shown in Fig. 25.14 is arranged so that successive input pulses reverse its state. In a shift register type the TR_2 output of the first triggers the TR_1 of the next.

26.6 Counters

Electronic counters are used in computers for counting the number of times an adding process has been carried out in multiplying, or subtracting in dividing. But they are also used much more widely; for example, counting articles passing on a conveyor belt, each article generating a pulse by interrupting a beam of light falling on a light-sensitive type of transistor. In high-precision frequency meters a counter is switched on for (say) exactly one second, during which the number of cycles is counted. Or by using a known frequency, an unknown time interval can be measured, also with great precision.

We are familiar with mechanical decimal counters, on car dashboards, for example. A mechanical binary counter would have only two symbols, 0 and 1, on each wheel. The first incoming 'pulse' would move it from 0 to 1; the second, back to 0 and at the same time the wheel would pulse the second wheel from 0 to 1. And so on. Obviously we can do the same thing electronically, using a chain of bistables of the Fig. 25.14 type in place of wheels. Fig. 26.6 shows in some detail how.

At *a* the circuits of the first two bistables in the series are shown in simplified form, omitting for clarity such practical details as the diodes in Fig. 25.14 needed to delete positive pulses and steer negative ones to the 'on' transistor. The waveforms at the points marked are shown at *b*. On the top line are the pulses to be counted.

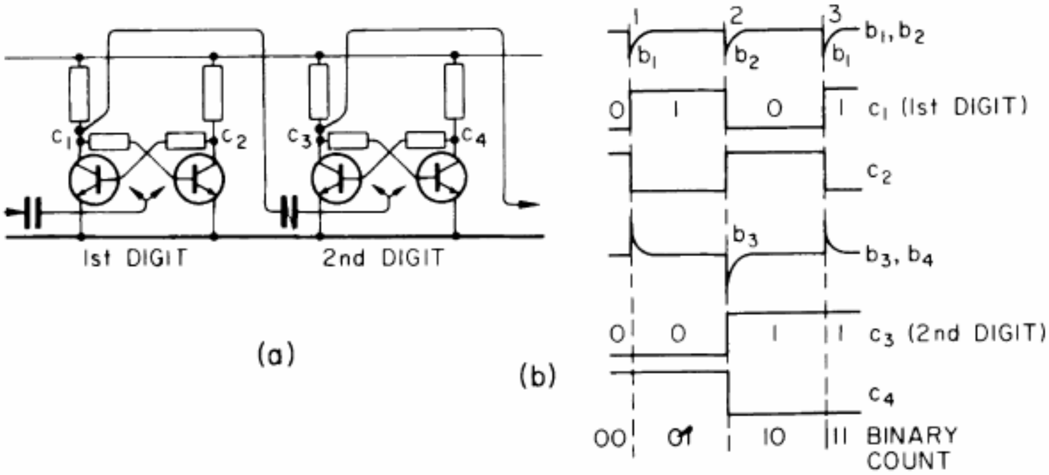


Fig. 26.6 (a) A two-digital place binary counter, using one bistable per place, with (b) waveforms during a count of 3 pulses

They need not be at regular intervals. 'Voltage positive' is taken to mean 1, so is read from the first collector in each bistable, but equally well it could be 'transistor on' or 'current flowing' and read from the second. The second bistable is triggered at the even-numbered input pulses and indicates the second digit. If each bistable has, say, a lamp to show whether its second half is on or off, and the lamps are placed in a row, right to left, the total of a series of pulses from one or more sources could be read off in binary notation, as shown on the last line in Fig. 26.6*b* for the first two. For convenience a binary-to-decimal converter could be interposed.

26.7 Stores

As mentioned earlier, digital computers require two types of store, one external to the computer proper to hold what might be called archive information, often with a very large number of items. Typical of the information held in such storage systems are payroll details of the staff of a large organisation. Magnetic tape is a suitable medium for a store of this type but the signals recorded on the tape are not, of course, like audio or video signals (which are analogue in nature) but are pulses, i.e., alternations between two signal levels which represent the binary digits 0 to 1. This is why punched paper tape and punched cards were used for such storage—a hole signifying 1 and the absence of a hole 0.

The capacity of a store could be expressed as the number of bits, i.e., binary digits it can hold, but the bit is a very small unit for mass storage systems and the capacity is usually expressed in bytes, a byte being a succession of bits (often eight in number) which is treated as a program can be erased and then replaced by a new program. Erasing is done by exposing the i.c. to ultra-violet light through windows in the package. This causes holes and electrons to combine, thus eliminating the stored charges.

26.8 Computer Languages

A digital computer responds only to inputs in binary form, and this applies to signal inputs on which the computer has to carry out the required operations and to the instructions required by the machine to enable it to carry out those operations. The instructions in binary form are known as 'machine language' and it is very longwinded. For example, more than a dozen separate instructions are needed to enable a computer to add two single-digit numbers. To speed up operation the instructions are therefore given in a shortened coded form known as a computer language, and one of the simplest of these is BASIC (Beginner's All-Purpose Symbolic Instruction Code). BASIC enables two seven-digit numbers to be added using only five instructions. It would be very convenient, of course, if computers could be programmed in ordinary English and some of the more recent so-called 'high-level' languages can be regarded as a move in that direction. Computers must therefore translate instructions into machine language, and the control unit (Fig 26.1) does this in addition to fetching any instructions required from the internal stores. The control unit also has a clock, i.e., a generator of regular pulses which are used to time the operations and to ensure that they are carried out in the right order.

26.9 Input and Output Devices

The input to a computer can be fed in via a keyboard resembling that of a typewriter, or from tape or disc. The output can be printed on paper or can be displayed in the form of words, numerals or diagrams on the screen of a Visual Display Unit (VDU), i.e., a cathode-ray tube which can also be used for displaying the input data. Domestic computers have a modulated-r.f. output which can be fed to a normal television receiver to display the computer output on the screen. Computers are now extensively employed in mechanical and electronic design and a new science known as Computer-aided Design (CAD) has evolved. For example the best way of interconnecting the individual components in a complex i.c. is now determined by use of the computer. Computers also play a large part in modern manufacturing methods and the techniques are collectively known as CAM (Computer-aided Manufacture).

26.10 Microprocessors and Microcomputers

A microprocessor consisting of an arithmetic and logic unit, control unit, a RAM and a ROM can be contained within a single i.c. a few millimetres square. Such i.cs are not costly and are now being increasingly used in applications where switching operations must be carried out in sequence and/or under the control of input signals or stored information. For example they can control switching in telephone exchanges and machine tools in factories. At home they control washing machines, make possible press-button telephone dialling, memorise and call favourite telephone numbers, and repeat any call number as often as desired, if not available the first time, by pressing one button.

So far, we have considered computers as individual sets of equipment, comparable with television sets. But to a rapidly increasing extent they are used in conjunction with other computers, sometimes in the same office, but often interconnected via telephone lines or even satellite transoceanic radio links. In such ways quite a modest domestic or office computer can avail itself of the resources of any of many more capable ones.

This is an example of the many ways, which in this book can be little more than mentioned, in which the 'electronic age' is coming into all our lives.

CHAPTER 27

Power Suppliers

27.1 The Power Required

In the preceding chapters the power supplies needed to make things work have been assumed. Let us now look at them in detail. Power for valves and cathode-ray tubes can be divided into three classes: cathode-heating, anode supplies, and grid bias.

So far as cathode heating is concerned, it is purely a matter of convenience that it is done electrically. In principle there is no reason why it could not be done by a bunsen burner or by focusing the sun on it with a lens. But there are overwhelming practical advantages in electric heating. Nearly always it is done, for convenience, by a.c. Cathode heating power has to be supplied all the time the valves are required for work, whether or not they are actually doing any.

Power supplied to the anode is more directly useful, in that a part of it is converted into the output of the valve. This is generally not true of power taken by auxiliary electrodes, so valves are designed to reduce that to a minimum.

Not all valves need any grid bias; and, of those that do, not all have to be provided with it from a special supply. And only in large sending valves is any appreciable power consumed, the reason being that with this exception care is taken to keep grid current as nearly as possible down to nil (Sec. 10.6). This restriction, desirable for most purposes, would unduly limit output from high-power valves.

Transistors have the great advantages of needing no heater power at all and of making more economical use of the power that is supplied. Often the sole source of power is a small dry battery, or (where available) a car battery. Larger, permanently-installed transistor equipment can more conveniently and economically be supplied (like most valve equipment) from the mains.

27.2 Batteries

A battery consists of a number of *cells*, and although a single cell is commonly called a battery it is no more correct to do so than to call a single gun a battery.

A cell consists of two different sorts of conducting plates or electrodes separated by an ‘exciting’ fluid or *electrolyte*. The e.m.f. depends solely on the materials used, not on their size, and cannot greatly exceed 2 V. The only way of obtaining higher voltage is to connect a number of cells in series to form a battery.

Where the need for compactness and long life (as in hearing aids and watches) justifies the greater cost, the mercury type of cell is preferred. The positive electrode is steel, the negative zinc, the electrolyte potassium hydroxide and the depolariser mercuric oxide. The voltage remains remarkably constant at about 1.2. These tend to be displaced by lithium cells (up to 3V) and zinc–air cells (1.65V).

The most commonly used secondary cell is the lead-acid type, in which the plates consist of lead frames or grids filled with different compounds of lead, and the electrolyte is dilute sulphuric acid. Accumulators have an e.m.f. of 2 V per cell and possess the great advantages of very low internal resistance and of being rechargeable.

Against this there are certain drawbacks. Lead is heavy. The acid is very corrosive, and liable to cause much damage if it leaks out or creeps up to the terminals. If the cell is allowed to stand for many weeks, even if unused, it becomes discharged, and if left in that condition it 'sulphates', that is to say, the plates become coated with lead sulphate which cannot readily be removed and which permanently reduced the number of ampere-hours the cell can yield on one charge. If the terminals are short-circuited, the resistance of the accumulator is so low that a very heavy current—possibly hundreds of amps—flows and is liable to cause permanent damage.

The size, weight and the presence of free acid makes the lead-acid battery unsuitable for some applications where a rechargeable battery would be useful, e.g. in portable television receivers. A better contender is an alternative type of secondary cell, available in sealed construction, which has electrodes of nickel and cadmium, and an electrolyte of potassium hydroxide. The e.m.f. is 1.2V, less than that of a lead-acid or Leclanché cell, so that more cells are needed to give a desired voltage, but the cells are small and light enough for use in portable equipment.

27.3 D.C. from A.C.

Power from the mains costs not more than about one fortieth that of power from dry batteries, so there is a strong incentive to use it except where very little power is needed or portability is essential. Almost everywhere mains power is in the form of a.c. Except for cathode heating, a.c. cannot be used as it is, because the requirement is for constant current or voltage in one direction. So we have to convert from a.c. to d.c. and therefore an essential part of a power unit is a *rectifier.*

If the a.c. is sinusoidal (as it usually nearly is) it can be represented as in Fig. 27.2*a*. The result of connecting a rectifier in series with it is to abolish half of every cycle, as shown as *b*. The average of this is only about one-third of the peak value. (Sec. 18.3) and in any case it would be useless in this form because it is not continuous. By using a *reservoir capacitor* (Sec. 18.6) this defect is removed and at the same time the average is brought nearly up to the peak voltage.

The conditions are not quite the same as in a detector, because the current drawn is usually much larger and the frequency much lower. Unless the reservoir capacitance is enormous, therefore, it has time to discharge appreciably between one cycle and the next, as in Fig. 27.2*c*. This is described as d.c. with a ripple, or unsmoothed d.c., which would cause a hum if used to feed a receiver. The next requirement, then, is a *smoothing circuit*, which is a low-pass filter (Sec. 18.10) designed to impede the a.c. ripple while allowing the d.c. to pass freely. If the filter is effective the result is as in Fig. 27.2*d*, practically indistinguishable from current supplied by a battery.

In some applications, especially those in which varying amounts of current are drawn and passing through the resistance of the rectifier and filter cause undesirable variations in voltage, there is a need for *stabilisation* (Sec. 27.9) to keep the voltage automatically constant.

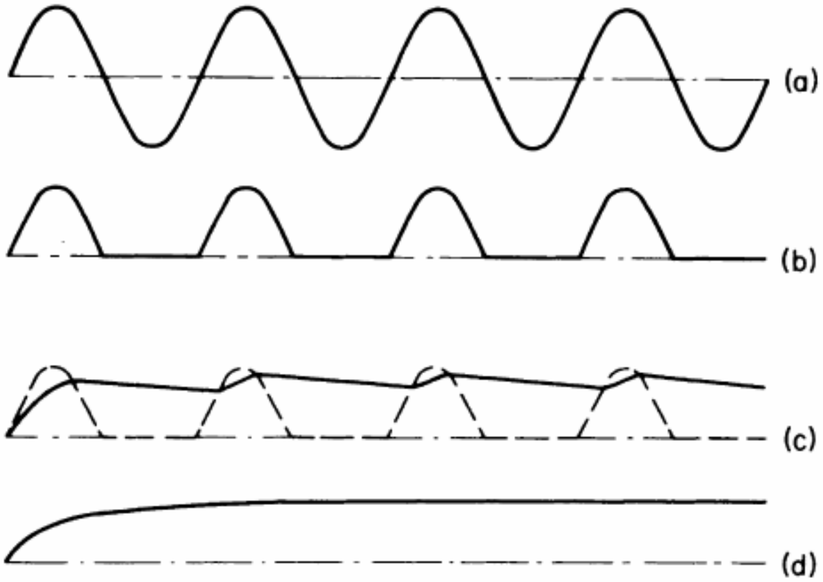


Fig. 27.2 (*a*) Original a.c.; (*b*) after rectification; (*c*) the effect of a reservoir capacitor; (*d*) after passing through a smoothing filter

Although in many domestic receivers, particularly television sets, the rectifier is connected straight to the mains, this is not suitable if the voltage required is much different from the mains voltage, and even when it is not the use of an intervening transformer is better practice because it improves safety by isolating the circuits from the mains, enabling them to be earthed.

27.4 Types of Rectifier

At one time vacuum diodes (Sec. 9.4) were used not only as signal voltage is taken off. This normally happens to a rectifier during half of every cycle. The usefulness of the gate is that it enables forward current to be prevented during any desired proportion of the other half-cycle and so enables the rectified output to be controlled, a facility used, for example, to control the brightness of mains-driven lamps.

27.5 Rectifier Circuits

There are various ways in which rectifiers can be connected in power unit circuits, and one of the simplest is indicated in Fig. 27.4*a*. A transformer T is shown for stepping the mains voltage up or down as required, and a single rectifier—the symbol denotes any type—is connected in series with it and a reservoir C_1. A simple filter circuit consisting of a choke L and a capacitor C_2 completes the apparatus.

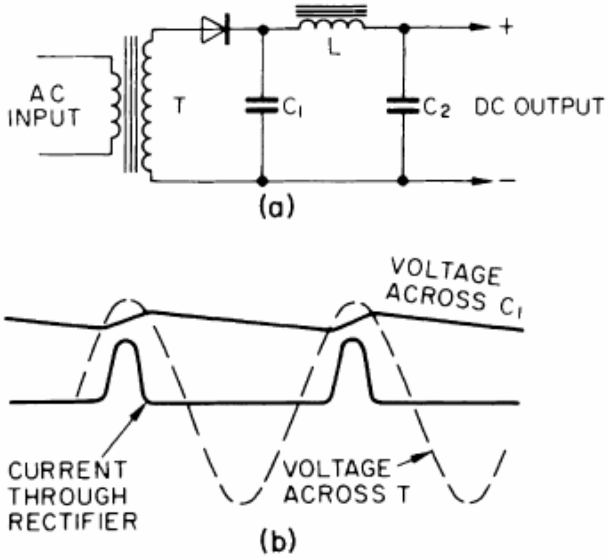


Fig. 27.4 The voltages and rectifier current in a simple half-wave rectifier power-unit circuit (*a*) are shown in diagram (*b*)

This is called a *half-wave* rectifier circuit, because only one-half of each a.c. cycle is utilised, the other being suppressed (Fig. 27.2*b*). If, as is desirable, the capacitance of C_1 is large enough for the voltage not to drop much between half-cycles, the proportion of each cycle during which the alternating voltage exceeds it is very small; and, as the recharging of C_1 has to take place during these brief moments, the rectifier is obliged to pass a current many times greater than that drawn off steadily at the output. If a large output current is needed, then the reservoir capacitance must be very large, the rectifier must be able to pass very heavy peak currents, and there are difficulties in designing the transformer to work under these conditions. So this circuit is used mainly for low-current high-voltage applications and twice the transformer secondary peak voltage. Compared with *d*, one output terminal is common to the source, which may be convenient for some purposes. As R_2C_2 is a half-wave rectifier this system suffers from the disadvantages of that type, as well as the losses in R_1, so is confined to high-voltage low-current power units.

(*f*) *Voltage Quadrupler.* By connecting a second type *e* circuit in parallel with the transformer, to give an equal voltage output of opposite polarity, the total output voltage tends towards four times the peak of the supply. The conversion of *e* to *f* is analogous to the conversion of *a* to *d*.

27.6 Filters

We have already used a simple filter circuit R_F and C_F in Fig. 18.10) to smooth out the unwanted r.f. left over from the rectification process in a detector. The same circuit is often used in power units, but the values of R_F and C_F are greatly different. The frequency of the ripple to be smoothed following a half-wave rectifier is usually 50 Hz, and 100 Hz with full-wave. For effective smoothing, the reactance of C_F at this frequency must be low compared with the resistance R_F. And because of the relatively heavy current in a power unit, R_F must itself be quite low if there is not to be an excessive wastage of volts.

For this reason an iron-cored choke coil is often substituted. It has to be carefully designed if the core is not to be saturated by the d.c. (Fig. 4.8) which greatly reduces the inductance and therefore the effectiveness of the coil for smoothing. A typical filter, LC_2, is seen in the half-wave rectifier circuit in Fig. 27.4.

Although a reservoir capacitor brings the rectified voltage almost up to the peak input voltage when no current is being taken, the voltage falls off fairly steeply with increase of load current. In technical language, the regulation is not very good. For purposes where it is important that the voltage should remain steady in spite of a fluctuating load current—for example, high-power Class B amplifiers—what is called a choke-input filter is used. In the circuit diagram this looks the same as, say, Fig. 27.4*a*, except that the reservoir capacitor is missing. This is not the only difference, however, for the choke has to be of a special kind, known as a swinging choke. Instead of being designed so that its inductance varies as little as possible, its inductance increases substantially as current falls, with the aim of maintaining a nearly constant voltage across it. The output voltage falls very rapidly up to, say, one tenth full load current (below which the current is not allowed to fall) and therefore is relatively steady. An advantage is that the current through the rectifier is relatively constant, instead of having peak values much greater than the mean as in Fig. 27.4*b*.

A choke having high reactance at low frequencies such as 100 Hz is large, heavy and costly. One way of avoiding the need for such components is to use a high ripple frequency. If, for example, the frequency can be raised to 25 kHz, the inductance can be reduced by a factor of 1/250 whilst maintaining the same reactance. Similarly the size of the smoothing capacitors can be reduced in the same ratio. A method of obtaining a high ripple frequency is described in Sec. 27.11. An alternative method of avoiding the use of a bulky choke at low ripple frequencies is to use a resistor for smoothing whenever possible, and for this reason electrolytic capacitors of very high capacitance—1 mF or more—are now customary. Even so, their impedance to the signal-frequency currents flowing through the power source may not be negligible, and a potential difference set up across this impedance by (say) the signal current in the output stage is passed on to all the other stages fed from the same supply and may seriously upset the working, possibly even causing continuous oscillation. We have, in fact, a form of feedback.

27.7 Decoupling

To obviate such undesirable conditions, *decoupling* is used. It is simply an individual filter for those stages liable to be seriously affected. As those stages are generally the preliminary ones, taking only a small proportion of the total current, a resistor having sufficient impedance at the frequencies concerned can be used. The loss in voltage may actually be desirable, as these stages are often required to run at lower voltages than the power output stage; in which case a single cheap component is made to serve the double purpose of a voltage-dropping resistance and (in conjunction with a capacitor) a decoupler. Of such is the essence of good commercial design.

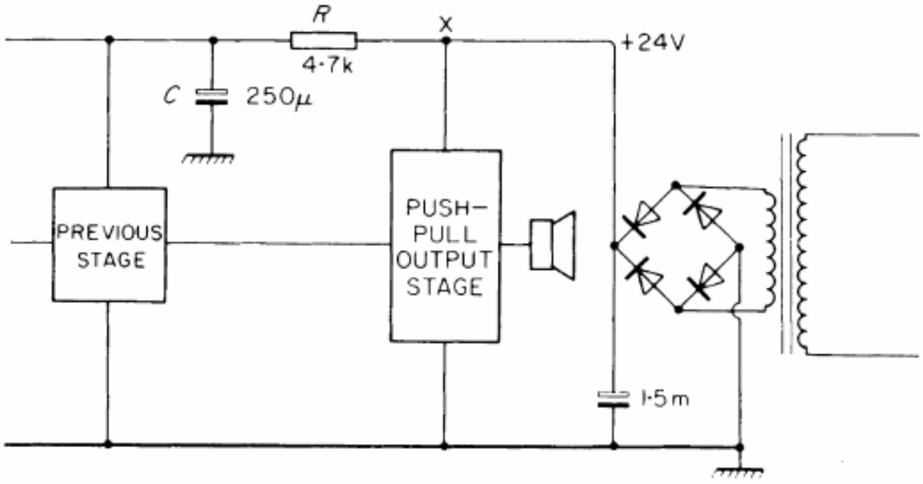


Fig. 27.6 Example of the use of decoupling

In Fig. 27.6, which is an outline circuit diagram of part of an a.f. amplifier, *R* and *C* are decoupling components. They serve as a filter, preventing signal voltages set up at point X by signal currents flowing through the power source from reaching the previous stage and then being amplified. In this example the output stage is fed direct from the reservoir capacitor, so there will be significant hum voltage at X. This may not matter to the output stage because it is push-pull, which tends to balance out hum, but it could be serious if it came in via the previous stage, so the extra smoothing given by *R* and *C* is needed for this reason too. Because of the comparatively small current flowing through *R* its resistance can be made very much larger and therefore more effective than could possibly be used in the main supply line. Signal currents in the stage fed by it find an easy path via *C* and so are kept out of other parts of the amplifier.

It is the extensive use of decoupling and filtering that makes multistage circuit diagrams look so complicated. But once their purpose has been grasped it is easy to sort them out from the main signal circuits. Note the use of a detached 'chassis' symbol at the foot of *C* to avoid drawing a confusing crossing wire to the base line. Considerable clarification of complicated circuits can be obtained in this way.

27.8 Bias Supplies

From time to time we have considered various methods of providing bias, and we now review them.

The most obvious way is by means of a battery or other source in series with the circuit connecting the two electrodes concerned, as in Figs. 10.6 and 11.2. A slight modification is the parallel-feed system (Fig. 13.4 and many others) where R_B is a resistance so high that it causes negligible loss of the signal applied via *C*. A choke coil is sometimes used instead of high resistance if the bias voltage must not be subject to drop due to current through it.

Obviously it would be possible to replace a bias battery by a mains power unit. Except for high-power equipment this is not done, because the power required is so small and there are more economical alternatives.

With valves and f.e.ts, which normally are biased opposite in polarity to the anode or drain, one of those alternatives is to insert a resistor as explained in Secs. 13.5 and 6. The required amount of bias is obtained by choice of resistance. To some extent this source of bias is self-adjusting; if an excessive current flows the bias voltage is correspondingly large and the rise in current is restrained. Also it is convenient for providing each stage with its own most suitable bias. Strictly it is cathode or source bias that is supplied, at the expense of anode or drain voltage, and this loss may sometimes need to be taken into account.

If only a simple resistor is used, it carries not only the steady (or z.f.) component of current but also the signal current, which produces a voltage in opposition to the input. We have, in fact, negative feedback (Sec. 19.15). If that is wanted, well and good; but if not, then something has to be done to provide a path of relatively negligible impedance for the signal currents, without disturbing the z.f. resistance. A by-pass capacitor, C_S in Fig. 13.8, serves the purpose; for r.f. a value of 0.1 μF or even less is enough, but for the lowest audio frequencies 50 μF or 100 μF may be needed.

Sometimes it is convenient to have a bias voltage that is equal to the peak value of the incoming signal. This is obtained by using the clamping principle (Sec. 25.8). With a.f. signals it is sometimes used for the stage immediately following a detector. Although a separate diode detector is commonly used, the gate-source or grid-cathode diode can be made to serve this purpose.

A very common application of this type of bias is for oscillators (Sec. 14.10); the fact that the bias voltage is nearly equal to the peak oscillation voltage means that it increases and so acts as a governor to prevent the amplitude of oscillation from increasing excessively. If the positive feedback in the oscillator is very large and the time constant equal to the duration of many cycles, the first burst of oscillation may generate such a large bias that the device so biased is cut off, so quenching the oscillation, which cannot resume until some of the charge has had time to leak off the grid. Oscillation then restarts, is again quenched, and so on. This phenomenon of intermittent oscillation is known as *squegging,* and is sometimes usefully empolyed. Elsewhere, it is a nuisance, which can be cured by reducing the positive feedback or the time constant of the biasing circuit or both. The blocking oscillator (Sec. 25.4) is sometimes considered as an extreme form of squegging.

27.9 Stabilisation

Often there is a need to maintain the terminal voltage of a power unit constant despite changes in the amount of current drawn. Sometimes the dual problem arises: to keep the output current constant regardless of the voltage. In the first case the desired result would be obtained if the internal resistance of the power unit was reduced to zero; in the second, if it was increased to infinity. We have seen (Sec. 19.17) that the tendency of negative feedback is to do one or other of these things, according to how it is applied, and in fact many stabilised power units can be regarded as negative-feedback

R_B which is chosen to pass a reasonable current through it, corresponds to the 'constant voltage' box in Fig. 27.7. The current to be kept constant, passing through R_E, provides the voltage V_E to be compared with V_Z. The difference between them is applied to TR. Because of the high mutual conductance of the transistor chosen for this role, a very small change in V_{BE}, caused by any change in I_E, brings into play a very steep correcting action. Changes in load resistance, R_L, cause corresponding changes in V_{EC}, but we know (from Fig. 11.7, for instance) that this has very little effect on I_C (and hence I_E) and such effect as it has is corrected as just explained. But of course there is a limit to the load resistance through which any chosen current can be passed. It occurs when the voltage across R_L nearly equals V_{CC} and the transistor is bottomed. So a fairly high value of V_{CC} (but not exceeding the rated maximum for the transistor) is desirable. The higher V_Z, the stronger the correcting action. The amount of current to be stabilised can be controlled by varying R_E.

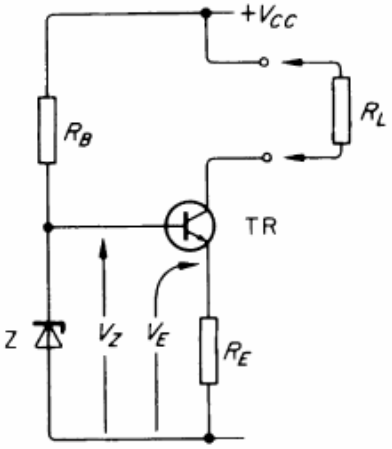


Fig. 27.9 Simple type of current stabiliser circuit

27.10 E.H.T.

Cathode-ray tubes in television and radar equipment need a voltage which is measured in kilovolts; 25 kV is a usual figure for colour television receivers. Fortunately the amount of current needed is quite small, of the order of 1 mA. A step-up mains transformer to develop such a voltage in the conventional way by a circuit such as Fig. 27.4 would be impracticably large, heavy and expensive, and also lethal. These objections are only partly removed by using the voltage quadrupler circuit (Fig. 27.5*f*). So, as mentioned in Sec. 23.10, use is made of the very high rate of change of current in the line-scanning transformer, seen in Fig. 23.20. The voltage developed in an inductor being proportional to the rate of current change, far fewer turns of wire are needed than with the very slow rate obtained with 50 Hz sinusoidal current for equal peak values.

capacitors in both left and right hand columns remain practically constant, but relative to the right-hand (output) column the left-hand one as a whole rises and falls in potential between —E and + E owing to the generator.

27.11 Switch-mode Power Suppliers

A useful development in power suppliers is the 'switch-mode' circuit in which the output from a mains rectifier feeds an astable pulse generator such as a multivibrator. This regularly chops the rectifier output to form an alternating voltage which is transformed, rectified and smoothed to give the required z.f. output voltage. This may seem an unnecessarily complicated method of obtaining a z.f. supply, but it has important advantages. As already pointed out (Sec. 27.6), if the multivibrator operates at a high frequency (e.g., 25 kHz) effective smoothing can be achieved with quite small inductors and capacitors. Moreover the output voltage of the supplier can be controlled by adjusting the mark/space ratio of the multivibrator (which determines the fraction of each multivibrator cycle during which the high-frequency rectifier conducts) and this can be used in an elegant method of stabilising the output voltage. A sample of the output voltage is compared with a standard constant voltage (as in Sec.27.9) and the difference, after ampliciation, is used to control the mark/space ratio. Power suppliers of this type are very efficient and compact, and are often used in television receivers.

27.12 Cathode Heating

As stated in Sec. 9.3, the cathode of a valve or c.r. tube may be either directly or indirectly heated. The power for either type can be obtained from a battery, but in practice directly heated valves are then used because they require only about one-tenth of the power needed by the indirectly heated types, and therefore a given battery will run them for much longer, or alternatively a much smaller battery can be used. These points are important, because battery power is so much more expensive than that drawn from the mains.

The vast majority of public electricity supplies are a.c., which can be stepped down to any convenient voltage for valve heating by means of a transformer. If an attempt were made to run directly heated valves in this way it would be unsuccessful, because the slight variation in temperature between peak and zero of the a.c. cycle, and also the variations of potential between grid and filament from the same cause, would produce variations in anode current, which would be amplified by the following stages and cause a loud low-pitched hum. A possible exception is a carefully arranged output stage, which of course is not followed by any amplification.

APPENDIX A

Algebraic Symbols

A.1 Letter Symbols

If a car had travelled 150 kilometres in 3 hours, we would know that its average speed was 50 kilometres per hour. How? The mental arithmetic could be written down like this:

$$150 \div 3 = 50$$

That is all right if we are concerned only with that one particular journey. If the same car, or another one, did 12 kilometres in a quarter of an hour, the arithmetic would have to be:

$$12 \div \tfrac{1}{4} = 12 \times 4 = 48$$

To let anyone know the speed of any car on any journey, it would be more than tedious to have to write out the figures for every possible case. All we need say is, 'To find the average speed in kilometres per hour, divide the number of kilometres travelled by the number of hours taken'.

What we actually would say would probably be briefer still: 'To get the average speed, divide the distance by the time'. Literally that is nonsense, because the only things that can be arithmetically divided or multiplied are numbers. But of course the words 'the number of' are (or ought to be!) understood. The other words—'kilometres per hour', 'kilometres', 'hours'—the units of measurement, as they are called, may also perhaps be taken for granted in such an easy case. In others, missing them out might lead one badly astray. Suppose the second journey had been specified in the alternative form of 12 kilometres in 15 minutes. Dividing one by the other would not give the right answer in kilometres per hours. It would, however, give the right number of kilometres per minute.

In electronics, as in other branches of physics and engineering, this matter of units is often less obvious, and always has to be kept in mind.

Our instruction, even in its shorter form, could be abbreviated by using mathematical symbols as we did with the numbers:

$$\text{Average speed} = \text{Distance travelled} \div \text{Time taken.}$$

Here we have a concise statement of general usefulness; note that it applies not only to cars but to aircraft, snails, bullets, space capsules, and everything else that moves.

Yet even this form of expression becomes tedious when many and complicated statements have to be presented. So for convenience we might write:

$$S = D \div T$$

or alternatively $$S = \frac{D}{T}$$

or, to suit the printer, $$S = D/T$$

A.2 What Letter Symbols Really Mean

This is the stage at which some people take fright, or become impatient. They say, 'How *can* you divide D by T? Dividing one letter by another doesn't *mean* anything. You have just said yourself that the only things that can be divided are numbers!' Quite so. It would be absurd to try to divide D by T. Those letters are there just to show what to do with the numbers when you know them. D, for example, has been used to stand for the number of kilometres travelled.

The only reason why the S, D and T were picked for this duty is that they help to remind one of the things they stand for. Except for that there is no reason why the same information should not have been written as:

$$x = \frac{y}{z}$$

or even

$$\alpha = \frac{\beta}{\gamma}$$

so long as we know what these symbols were intended to mean. As there are only 26 letters in our alphabet, and fewer still in the Greek, we cannot allocate any one of them permanently to mean 'average speed in kilometres per hour'. x, in particular, is notoriously capable of meaning absolutely anything. Yet to write out the exact meaning every time would defeat the whole purpose of using the letters. How, then, does one know the meaning?

Well, some meanings have been fixed by international agreement. There is one symbol that means the same thing every time, not only in electronics but in all the sciences—the Greek letter π (read as 'pie'). It is a particularly good example of abbreviation because it stands for a number that would take eternity to write out in full—the ratio of the circumference of a circle to its diameter, which begins: 3.1415926535 . . . . . (The first three or four decimal places give enough accuracy for most purposes.)

Then there is a much larger group of symbols which have been given meanings that hold good throughout a limited field such as electrical engineering, or one of its subdivisions such as electronics, but are liable to mean something different in, say, astronomy or hydraulics. These have to be learnt by anyone going in for the subject seriously.

Lastly, there are symbols that one uses for all the things that are not in a standard list. Here we are free to choose our own; but there are some rules it is wise to observe. It is common sense to steer clear as far as possible from symbols that already have established meanings. The important thing is to state the meaning when first using the symbol. It can be assumed to bear this meaning to the end of the particular occasion for which it was attached; after that, the label is taken off and the symbol thrown back into the common stock, ready for use on another occasion, perhaps with a different label—provided it is not likely to be confused with the first.

Sometimes a single symbol might have any of several different meanings, and one has to decide which it bears in that particular context. The Greek letter μ (pronounced 'mew') is an example that occurs in this book. When the subject is a radio valve, it can be assumed to mean 'amplification factor'. But if iron cores for transformers are being discussed, μ should be read as 'permeability'. And if it comes before an upright letter it is an abbreviation for 'micro-', meaning 'one millionth of'.

A.3 Some Other Uses of Symbols

The last of these three meanings of μ is a different kind of meaning altogether from those we have been discussing. Until then we had been considering symbols as abbreviations for quantities of certain specified things, such as speed in kilometres per hour. But there are several other uses to which they are put.

Instead of looking on '$S=D/T$' as an instruction for calculating the speed, we can regard it as a statement showing the relationship to one another of the three quantities, speed, time and distance. Such a statement, always employing the 'equals' sign, is called by mathematicians an *equation*. From this point of view S is no more important than T or D, and it is merely incidental that the equation was written in such a form as to give instructions for finding S rather than for finding either of the other two. We are entitled to apply the usual rules of arithmetic in order to put the statement into whatever form may be most convenient when we come to substitute the numbers for which the letters stand.

For example, we might want to be able to calculate the time taken on a journey, knowing the distance and average speed. We can divide or multiply both 'sides' of an equation by any number (known, or temporarily represented by a letter) without upsetting their equality. If we multiply both sides of $S = D/T$ by T we get $ST = D$ (ST being the recognised abbreviation for $S \times T$). Dividing both sides of this new form of the equation by S we get $T = D/S$. Our equation is now in the form of an instruction to divide the number of kilometres by the speed in kilometres per hour (e.g., 150 kilometres at 50 km/h takes 3 hours).

When you see books on electronics (or any other technical subject) with pages covered almost entirely with mathematical symbols, you can take it that instead of explaining in words how their conclusions are reached, the authors are doing it more compactly in symbols. It is because such pages are concentrated essense, rather than that the meanings of the symbols themselves are hard to learn, that makes them difficult. The procedure is to express the known or assumed facts in the form of equations, and then to combine or manipulate these equations according to the established rules in order to draw some useful or interesting conclusions from those facts.

This book, being an elementary one, explains things in words, and only uses symbols for expressing the important facts or conclusions in concise form, or for rearranging them.

A.4 Abbreviations

Another use for symbols is for abbreviation pure and simple. We have already used one without explanation—km/h—because it is known that this means 'kilometres per hour'. The kilometre-per-hour is a unit of speed. '£' is a familiar abbreviation for the pound sterling. The unit of electrical pressure is the volt, denoted by the abbreviation V. Sometimes it is necessary to specify very small voltages, such as 5 millionths of a volt. That could be written 0.000005 V. But a more convenient abbreviation is 5 μV, read as '5 microvolts'. The list in Appendix D gives the abbreviations commonly used in electronics.

Still another use for letters is to point out details on a diagram. R stands for electrical resistance, but one has to guess whether it is intended to mean resistance in general, as a property of conductors, or the numerical value of resistance in an equation, or the particular resistance marked R in a diagram. Often it may combine these meanings, being understood to mean 'the numerical value of the resistance marked R in Fig. So-and-so'.

Attaching the right meanings to symbols probably sounds dreadfully difficult and confusing. So do the rules of a new game. The only way to defeat the difficulties is to start playing the game.

Here are a few more hints about symbols.

A.5 How Numbers Are Used

One way of making the limited stock of letters go farther is to use different kinds of type. It has become standard practice to distinguish symbols for physical quantities by italic (sloping) letters, leaving the roman (upright) ones for abbreviations; for example, 'V' denotes an unspecified amount of electrical potential difference, reckoned in volts, for which the standard abbreviation is 'V'. 'R' is the resistance of a resistor R. And so on.

Another way of making letters go farther is to number them. To prevent the numbers from being treated as separate things they are written small near the foot of the letter ('subscript'). For example, if we want to refer to several different resistances we can mark them R_1, R_2, R_3, etc. Sometimes a modification of a thing denoted by one symbol is distinguished by a tick or dash; A might stand for the amplification of a receiver when used normally, and A' when modified in some way.

But on no account must *numbers* be used 'superscript' for this purpose, because that already has a standard meaning. 5^2 (read as '5 squared') means 5×5; 5^3 ('5 cubed') means $5 \times 5 \times 5$; 5^4 ('5 to the 4th power') means $5 \times 5 \times 5 \times 5$; x^2 means the number represented by x multiplied by the same number.

x^{-2} seems nonsensical according to the rule just illustrated. It has been agreed, however, to make it mean $1/x^2$ (called the *reciprocal* of x squared). The point of this appears most clearly when these superscripts, called *indices* (singular: *index*), are applied to the number 10. $10^4 = 10000$, $10^3 = 1000$, $10^2 = 100$, $10^1 = 10$, $10^0 = 1$, $10^{-1} = 0.1$ $10^{-2} = 0.01$, and so on. The rule is that the power of 10 indicates the number of places the decimal point has to be away from 1; with a positive index it is on the right; negative, on the left. Advantage is taken of this to abbreviate very large or very small numbers. It is easier to see that 0.0000000026 is 2.6 thousand-millionths if it is written 2.6×10^{-9}. Likewise 2.6×10^{-12} is briefer and clearer than 2600000000000.

$x^{1/2}$ also needs explanation. It is read as 'the square root of x', and is often denoted by $\sqrt{x}$. It signifies the number which, when multiplied by the same number, gives x. In symbols, $(\sqrt{x})^2 = x$.

Note that $10^6 \times 10^6 = 10^{12}$, and that $10^{12} \times 10^{-3} = 10^{12}/10^3 = 10^9$. In short, multiply by adding indices, and divide by subtracting them. This idea is very important in connection with decibels (Sec. 19.2).

APPENDIX B

Graphs

B.1 What is a Graph?

Most of us when we were in the growing stage used to stand bolt upright against the edge of a door to have our height marked up. The succession of marks did not convey very much when reviewed afterwards unless the dates were marked too. Even then one had to look closely to read the dates, and the progress of growth was difficult to visualise. Nor would it have been a great help to have presented the information in the form of a table with two columns—Height and Date.

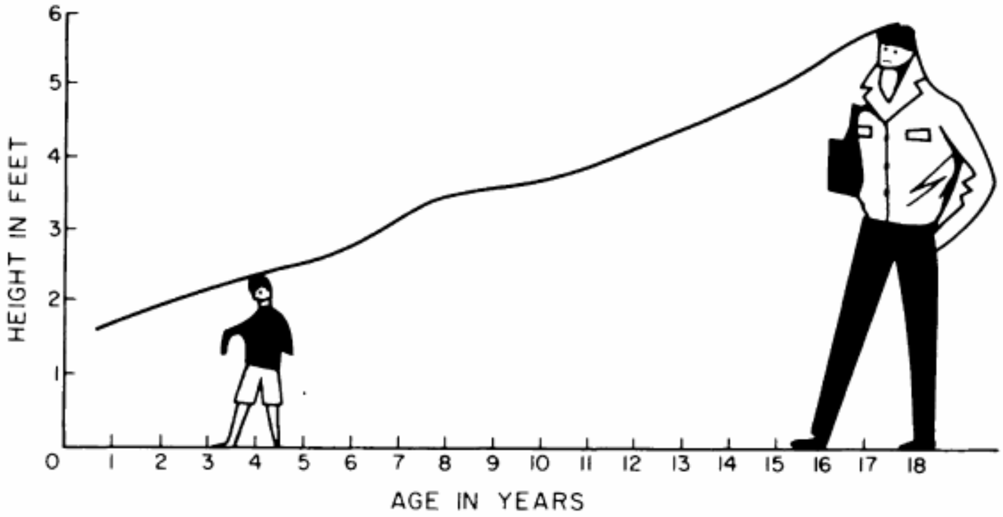


Fig. B.1 Simple (but inconvenient) system for automatic graph plotting of human growth. The equipment consists of a conveyor belt moving at the rate of one foot per year, a pencil, and a white wall

But, disregarding certain technical difficulties, imagine that the growing boy had been attached to a conveyor belt which moved him horizontally along a wall at a steady rate of, say, one foot per year, and that a pencil fixed to the top of his head had been tracing a line on the wall (Fig. B.1). If he had not been growing at all, this line would, of course, be straight and horizontal. If he had been growing at a uniform rate, it would be straight but sloping upwards. A variable rate of growth would be shown by a line of varying slope.

In this way the progress over a period of years could be visualised by a glance at the wall.

B.2 Scales

To make the information more definite, a mark could have been made along a horizontal line each New Year's Day and the number of the year written against it. In more technical terms this would be a time scale. The advantage of making the belt move at uniform speed is that times intermediate between those actually marked can be identified by measuring off a proportionate *distance.* If one foot represented one year, the height at, say, the end of May in any year could be found by noting the height of the pencil line 5 inches beyond the mark indicating the start of that year.

Similarly a scale of height could be marked anywhere in a vertical direction. It happens in this case that height would be represented by an equal height. But if the graph were reproduced in a book, as here, although height would still represent height it would have to do so on a reduced scale, say half an inch to a foot.

To guide the eye from any point on the information line to the two scales, it is usual to plot graphs on paper printed with horizontal and vertical lines so close together that a pair of them is sure to come near enough to the selected point for any little less or more to be judged. The position of the scales is not a vital matter, but unless there is a reason for doing otherwise they are marked along the lines where the other quantity is zero. For instance, the time scale would be placed where height is nil; i.e., along the foot of the wall. If, however, the height scale were erected at the start of the year 1 A.D. there would be over three-quarters of a kilometre of blank wall between it and the pencil line. To avoid such inconvenience we can use a 'false zero', making the scale start at or slightly below the first figure in which we are interested. A sensible way of doing it in this example would be to reckon from the time the boy was born, as in Fig. B.1. The point that is zero on both scales is called the *origin*.

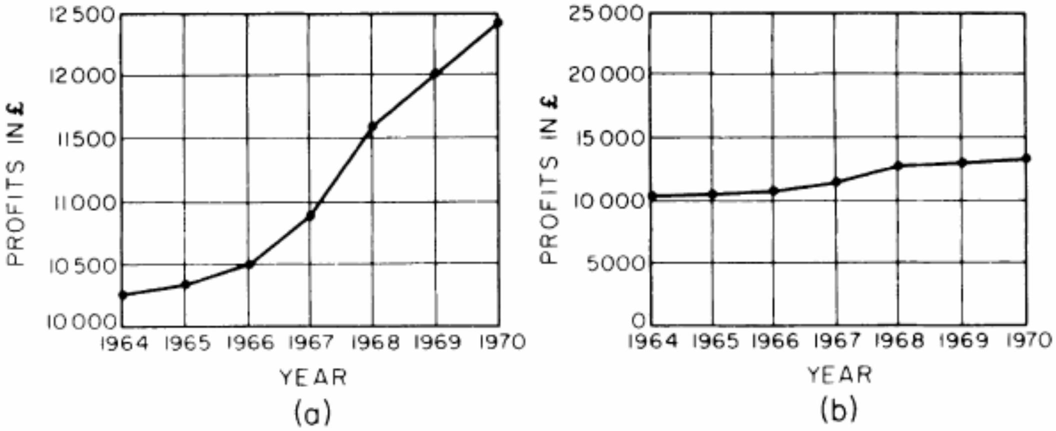


Fig. B.2 (*a*) Typical financial graph arranged to make the maximum impression consistent with strict truthfulness. (*b*) The same information presented in a slightly different manner

False zeros are sometimes used in a slightly shady manner to give a wrong impression. Fig. B.2*a* shows the sort of graph that might appear in a company-promoting prospectus. The word 'Profit' would, of course, be in a big type, and the figures indistinctly, so that the curve would seem to indicate a sensational growth in profits. Plotted without a false zero, as in Fig. B.2*b*, it looks much less impressive. Provided that it is clearly admitted, however, a false zero is useful for enabling the significant part of the scale to be expanded and so read more precisely.

Fig. B.2, by the way, is not a true graph of the sort a mathematician would have anything to do with. Profits are declared annually in lumps, and the lines joining the dots that mark each year's result have no meaning whatever but are there merely to guide the eye from one to another. In a true graph a continuous variation is shown, and the dots are so close together as to form a line. This line is technically a *curve*, whether it is what is commonly understood as a curve or is as straight as the proverbial bee-line.

B.3 What a 'Curve' Signifies

Every point on the curve in Fig. B.1 represents the height of the boy at a certain definite time (or, put in another way, the time at which the boy reached a certain height). Since it is possible to distinguish a great number of points along even a small graph, and each point is equivalent to saying 'When the time was T years, the height was H feet', a graph is not only a very clear way of presenting information, but a very economical one.

Here we have our old friend T again, meaning time, but now in years instead of hours. It is being used to stand for the number of years, as measured along the time scale, represented by *any* point on the curve. Since no particular point is specified, we cannot tell how many the corresponding height may be, so have to denote it by H. But directly T is specified by a definite number, the number or *value* of H can be found from the curve. And vice versa.

In its earlier role (Appendix A, Sec. A1) T took part in an equation with two other quantities, denoted by S and D. The equation expressed the relationship between these three quantities. We now see that a graph is a method of showing the relationship between two quantities. It is particularly useful for quantities whose relationship is too complicated or irregular to express as an equation—the height and age of a boy, for instance.

B.4 Three-dimensional Graphs

Can a graph deal with three quantities? There are three ways in which it can, none of them entirely satisfactory.

Each point plotted in *a* corresponds to an *S* curve in *b*, and when joined up they give a new curve (really curved this time!). This process has been demonstrated for only one value of *D*, but of course it could be repeated to show the time/speed relationship for other distances. Fig. B.3*b* gives numerical expression to the well-known fact that the greater the speed the shorter is the time to travel a specified distance.

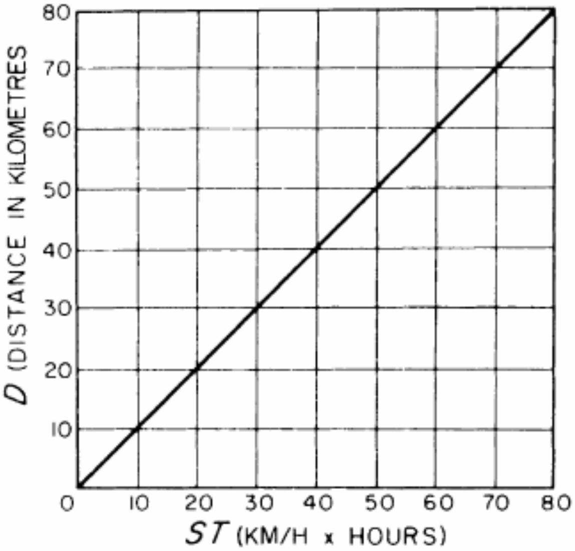


Fig. B.4 A third method of graphing three variables is to combine two of them; in this example the product of *S* and *T* is used as one of the two variables in the graph

Graphs of this kind are important in many branches of electronics; for example, with transistors, whose relationships cannot be accurately expressed as equations.

Lastly, two (or more) variables can be grouped so as to reduce the total to two. In our $D = ST$, D can be plotted against ST, which in mathematical language is called the product of S and T. The resulting very simple graph is shown as Fig. B.4. A point on the *ST* scale, say 60, shows the distance travelled in 2 hours at 30 kilometres per hour, 3 hours at 20 kilometres per hour, etc. No one would bother to draw a graph for such a simple equation, but this is quite a useful technique for more complicated ones. Fig. 20.3 is an example.

B.5 Significance of Slope

Even from simple two-dimensional graphs like Fig. B.1 it may be possible to extract information that is worth replotting separately. It is clear that a third quantity is involved—*rate of growth*. This is not an independent variable; it is determined by the other two. For it is nothing else than rate-of-change-of-height. As we saw at the start, when the curve slopes upward steeply, growth is rapid; when the curve flattens out, it means that growth has ceased; and if the curve started to slope downwards—a negative gradient—it would indicate negative growth. So it would be possible to plot growth in, say, inches per year, against age in years. This idea of rate-of-change, indicated on a graph by slope, is the essence of no less a subject than the differential calculus, which is much easier than it is often made out to be. If you are anxious to get a clearer view of any electrical subject it would be well worth while to read at least the first few chapters of an elementary book on the differential calculus, such as S. P. Thompson's classic 'Calculus Made Easy' (Macmillan).

B.6 Non-Uniform Scales

Finally, although uniform scales for graphs are much the easiest to read, there may quite often be reasons which justify some other sort of scale. The most important, and the only one we need consider here, is the *logarithmic* scale. In this, the numbers are so spaced out that a given distance measured anywhere along the scale represents, not a certain *addition,* but a certain *ratio* or *multiplication.* Musicians use such a scale (whether they know it or not); their intervals correspond to ratios. For example, raising a musical note by an octave means doubling the frequency of vibration, no matter whereabouts on the scale it occurs. Users of slide-rules were familiar with the same sort of scale. Readers who are neither musicians nor slide-rule pushers can see the difference by looking at Fig. B.5, where *a* is an ordinary uniform (or *linear*) scale and *b* is logarithmic. If we start with *a* and note any two scale readings 1½ inches apart we find that they always *differ* by 10. Applying the same test to *b* we find that the difference may be large or small, depending on whether the inch and a half is near the top or bottom. But the larger reading is always 10 times the smaller.

One advantage of this is that it enables very large and very small readings to be shown clearly on the same graph. Another is that with some quantities (such as musical pitch) the ratio is more significant than the numerical interval. Fig. 20.7 is an example of logarithmic scales.

APPENDIX C

Alternative Technical Terms

The reader of books and articles on radio is liable to be confused by the use of different terms to mean the same thing. In the following list the first alternatives to be mentioned are those preferred in this book. The associated terms are not necessarily exact equivalents. Terms distinctively American are printed in italics. The section numbers refer to where the terms are defined or explained.

Accumulator— Secondary battery	27.2
Amplification—Gain	11.5
Anode—*Plate*	9.4
Anode a.c. resistance—Anode incremental resistance—Valve impedance—*Plate impedance*	10.4
Anode battery—H.T.—*'B' battery*	9.4
Antenna—Aerial	17.7
Astable—Multivibrator	25.5
Atmospherics—Strays—X's—*Static*	17.13
Audio frequency (a.f.)—Low frequency (l.f.)—Speech frequency—Voice frequency	1.4
Automatic gain control (a.g.c.)—Automatic volume control (a.v.c)	22.11
Bistable—Eccles-Jordan relay	25.6
Blumlein integrator—Miller integrator	25.2
Bottom bend—Cut-off	11.9
Bottoming—Saturation	11.9
Capacitance—Capacity	3.2
Capacitor—Condenser	3.3
Characteristic resistance—Characteristic impedance—Surge impedance	16.3
Coaxial—Concentric	16.1
Common-anode valve—Cathode follower	12.9
Common cathode (or grid, anode, emitter, base, collector, source, gate, drain)—Earthed cathode (etc.)—*Grounded cathode* (etc.)	12.9
Common-collector transistor—Emitter follower	12.9
Compound pair (of transistors)—Darlington pair—Super-alpha pair—Double emitter follower	19.14
Detection—Rectification—Demodulation (British usage originally reserved this term for a different phenomenon; Secs. 18.1, 22.10)	18.1
Dielectric—Insulating material	3.3
Dynamic resistance—*Antiresonant impedance*	8.10

APPENDIX D

Symbols and Abbreviations

General abbreviations

a.c.	alternating current
a.f.	audio frequency
a.g.c.	automatic gain control
a.m.	amplitude modulation
a.v.	alternating voltage
c.r.t.	cathode-ray tube
c.s.r.	controlled silicon rectifier
d.c.	direct current
d.l.	diode logic
d.s.b.	double sideband
d.v.	direct voltage
e.h.t.	extra-high tension
e.m.f.	electromotive force
EPROM	erasable, programmable, read-only memory
f.e.t.	field-effect transistor
f.m.	frequency modulation
h.f.	high frequency
h.t.	high tension
i.c.	integrated circuit
i.f.	intermediate frequency
igfet	insulated-gate field-effect transistor
i.r.	infra red
i.s.b.	independent sideband
jugfet	junction-gate field-effect transistor
l.f.	low frequency
l.i.c.	linear integrated circuit
l.s.	loudspeaker
p.c.b.	printed-circuit board
p.d.	potential difference
p.d.a.	post-deflection acceleration
p.i.l.	precision in line
p.i.v.	peak inverse voltage
PROM	programmable read-only memory
p.c.b.	printed circuit board
RAM	random access memory
r.f.	radio frequency
r.g.b.	red, green, blue
r.m.s.	root-mean-square
ROM	read-only memory
s.a.w.	surface acoustic wave
s.c.r.	silicon controlled rectifier
s.s.b.	single sideband
s.w.r.	standing wave ratio

APPENDIX E

Decibel Table

The decibel figures are in the centre column: figures to the left represent decibel loss, and those to the right decibel gain. The voltage and current figures are given on the assumption that there is no change in impedance.

Voltage or current ratio	*Power ratio*	— *dB* +	*Voltage or current ratio*	*Power ratio*
1.00	1.000	0	1.000	1.000
0.989	0.977	0.1	1.012	1.022
0.977	0.955	0.2	1.023	1.047
0.966	0.933	0.3	1.035	1.072
0.955	0.912	0.4	1.047	1.096
0.944	0.891	0.5	1.059	1.122
0.933	0.871	0.6	1.072	1.148
0.912	0.832	0.8	1.096	1.202
0.891	0.794	1.0	1.122	1.259
0.841	0.708	1.5	1.189	1.413
0.794	0.631	2.0	1.259	1.585
0.750	0.562	2.5	1.334	1.778
0.708	0.501	3.0	1.413	1.995
0.668	0.447	3.5	1.496	2.239
0.631	0.398	4.0	1.585	2.512
0.596	0.355	4.5	1.679	2.818
0.562	0.316	5.0	1.778	3.162
0.501	0.251	6.0	1.995	3.981
0.447	0.200	7.0	2.239	5.012
0.398	0.159	8.0	2.512	6.310
0.355	0.126	9.0	2.818	7.943
0.316	0.100	10	3.162	10.00
0.282	0.0794	11	3.55	12.6
0.251	0.0631	12	3.98	15.9
0.224	0.0501	13	4.47	20.0
0.200	0.0398	14	5.01	25.1
0.178	0.0316	15	5.62	31.6
0.159	0.0251	16	6.31	39.8
0.126	0.0159	18	7.94	63.1
0.100	0.0100	20	10.00	100.0
3.16×10^{-2}	10^{-3}	30	3.16×10	10^{3}
10^{-2}	10^{-4}	40	10^{2}	10^{4}
3.16×10^{-3}	10^{-5}	50	3.16×10^{2}	10^{5}
10^{-3}	10^{-6}	60	10^{3}	10^{6}
3.16×10^{-4}	10^{-7}	70	3.16×10^{3}	10^{7}
10^{-4}	10^{-8}	80	10^{4}	10^{8}
3.16×10^{-5}	10^{-9}	90	3.16×10^{4}	10^{9}
10^{-5}	10^{-10}	100	10^{5}	10^{10}
3.16×10^{-6}	10^{-11}	110	3.16×10^{5}	10^{11}
10^{-6}	10^{-12}	120	10^{6}	10^{12}

Index

Abbreviations, 523, 534–537,
A.C. power supplies, 504–510
Acceptor circuit, 123
Acceptor impurity, 139
Accumulator, 53, 503–504
Active device, 172
Adder, 494–496
Admittance, 95
Aerial. *See* Antenna
After-glow, 416, 446
Air traffic control, 446
Alternating current, 69–82
  addition and subtraction, 78
  frequencies of, 70
  meters, 73
Alternating voltage, 69
  addition and subtraction, 69
Ammeter, 23, 39
Ampere, 22, 25, 41
Amplification, 173, 175
  a.f., 321–342
  calculation, 179
  r.f., 387–414
  v.f., 449–460
  voltage, 156, 175, 180
Amplification factor, 152, 179, 522
  current, 164
  voltage, 152
Amplifier, 15
  a.f., 333–342
  aperiodic, 387
  balanced, 462–463
  cascode, 401
  Class A, 183, 334
  Class B, 183, 246, 335–342, 353
  Class C, 183, 246
  d.c., 454, 462, 478, 484
  direct-coupled, 454, 462, 478, 484
  differential, 462
  i.f., 396–399
  inverter, 462–463, 489
  long-tailed pair, 462–463
  microwave, 243, 413–414
  non-inverting, 462–463
  operational, 461–464, 468, 471
  pulse, 449–460
  r.f., 321, 399–403, 414
Amplifier *(cont.)*
  u.h.f., 402
  v.f., 449–460
  voltage, 156, 175
  wide-band, 449–460
Amplitude, 3
Amplitude modulation, 244–247, 251–254, 374, 375, 377, 378
Analogue computer, 462, 488
AND gate, 485
Anode, 131
Anode a.c. resistance, 154
Anode conductance, 154
Anode power supply, 502
Antenna, 281–287, 290–296
  coupling, 409–413
  dipole, 281–287, 290–293
  directional characteristics, 282
  director, 284
  effective height, 294
  folded dipole, 293
  frame, 295
  gain, 286
  horn, 295
  inductor, 295
  input circuit, 272
  loop, 295
  Marconi or quarter-wave, 290–291
  microwave, 294–295
  parasitic, 284
  reflector, 284
  slot, 285–286
  tuning, 293
  Yagi, 285, 293
Antinode, 271
Aperiodic coupling, 387
Arithmetic and logic unit, 490
Aspect ratio, 425
Astable system, 472
Atomic number, 134
Atoms, 18, 20
Attenuation distortion, 321
Audio frequencies, 6, 69, 107, 174, 321–342
Automatic frequency control, 357
Automatic gain control, 407–409, 482
Auto-transformer, 109, 229, 230

Gain *(cont.)*
  current, 198, 201, 204, 208
  voltage, 176, 197, 198, 201, 204, 208, 211
Gain/frequency distortion, 323–324, 331
Ganged tuning, 382, 383, 399, 403, 412
Gate
f.e.t., 165
  linear, 483–484
  logic, 484–486
Gate bias, 220–222
Gate insulation, 168
Gated a.g.c., 409
Generator, 21, 42, 70
  equivalent, 176–179, 183, 189, 236, 388
  pulse, 472–478
  sine-wave, 229–237
Germanium, 134, 136, 138, 141, 145
Germanium diode, 310
Germanium transistor, 161, 217
Graphs, 525–531
  three-dimensional, 527–529
Greek letters, 535
Grid, 151
Grid bias, 156, 502
Ground wave, 288–289
Gun, electron, 415

h parameters, 196, 199, 391
Half-adder, 495, 496
Harmonic distortion, 325–327, 343–345
Harmonics, 325–327, 385, 450–451
Hartley oscillator, 229, 230, 231, 319
Hartley's law, 439, 453
Heater, 130
Heating effect of electricity, 39
Heising modulation, 247
Henry, 60
Hertz, H., 11, 13, 14, 127
Heterodyning, 374
High tension, 131
Holes, 135–137
Horn antenna, 295
Hue, 440
Hum, 505
Hybrid-π equivalent circuit, 391–392
Hysteresis, 124

Igfet, 166
Image interference, 384
Impedance, 92, 95
  matching, 268–269, 270, 412
  transformation, 107
Impurity conduction, 138–141
Indices, 524
Inductance, 56–68, 223, 457–458
  analysis, 61
  in a.c. circuits, 97–109
  in parallel with capacitance, 117
  in parallel with capacitance and resistance, 125–127
  in parallel with resistance, 102, 108
  in series and in parallel, 99
  in series with capacitance, 110
  in series with capacitance and resistance, 111–113
  in series with resistance, 101
  leakage, 106
  mutual, 67, 68, 99, 367–369
  self-, 60, 99, 275
Induction, 59
Induction field, 279
Inductive circuit, electric current in, 64
  power in, 66–67
Inductor, 61, 383
Inductor antenna, 295
Input characteristic, 161
Input circuit, 171
Input resistance, 193–196, 198, 201, 204, 206, 349–351
Input signals, 173
Instruments for measuring current, 39, 40
Insulators, 21
Integrated circuits, 317, 319, 464, 476, 487, 489, 491
Integrating circuit, 465–469
Integration, 489
Intercarrier, 431
Interference, 369, 384–386
  second-channel or image, 384
Interlacing, 426
Intermediate frequency, 373, 382
Intermodulation, 327–328
Internal resistance, 38
Intrinsic conduction, 137, 138, 142, 149
Inversion of signal, 176
Ionisation, 133
Ions, 20
Iron core, 123–124
  in transformer, 68, 103

Johnson noise, 355
Joule, 54
Jugfet, 166

Keying, frequency shift, 249
Kinetic energy, 43, 67

Saturation, colour, 440
  current, 132, 183
Sawtooth waveform, 422
Scales, 526–527, 531
Scanning, 9, 10, 250, 426
Scanning circuits, 435–439
Schmitt trigger, 479
Screening, principle of electric, 399–401
  principle of magnetic, 403–405
Secondary cell, 503
Secondary emission, 416
Second-channel interference, 384
See-saw circuit, 462
Selectivity, 115, 255, 358–372, 401, 412
  and Q, 358–367
Selectivity curve, 361–371
Self-inductance, 60, 99, 275
Self-oscillating mixer, 379
Semiconductor, 21, 129, 130, 134
  i-type, 142
  metal-oxide-semiconductor, 166
  metal-oxide-silicon, 166
  n-type, 140, 149
  p-type, 140, 149
Semiconductor devices, 129
Semiconductor diode, 144, 235, 507
Sender, 7, 15, 242, 244, 246, 248–251, 258, 271, 286, 291, 292, 357, 385, 444
Sensitivity, 358
Series connection, 29, 30, 31, 34, 38, 44, 88, 91, 99, 101, 110, 111, 126
Series-parallel combinations, 31
Series resistance of a tuned circuit, 358
Shadow-mask picture tube, 444
Shock excitation, 226
Shelf life, 503
Shift register, 497
Short circuit, 37
Short-circuited-output current gain, 393
Shot effect, 355
Shunt, 41
Shunt-injected negative feedback, 349
Shunting, 29
SI units, 34, 45, 48, 60, 62
Sidebands, 251–256, 360, 361–372, 375, 429
  definition, 254
Siemens, 35
σ, 35
Sign convention, transistor equivalent generator, 190–191
Signal/noise ratio, 355, 381, 399, 401
Signal, 173
  input, 173
  inversion, 176
Silicon, 134, 135, 138, 141, 145
Silicon diode, 146
Silicon transistor, 160–164
Sine wave, 70, 76, 325
Single-sideband system, 256
Sinusoidal modulation, 244–247, 251–255
Skin-effect, 124
Skip distance, 288
Slope, 153, 529–530
Slot antenna, 285–286
Smoothing, 505, 510–511
Solid-state. *See* semiconductor
Sound generator, 5
Sound waves, 1–3, 12, 13, 69, 249, 332
Source, of an f.e.t., 165
  of a signal, 171–172
Source follower, 210
Space charge, 131, 132, 143, 145, 151, 163, 391
Space-charge-limited current, 132
Speed-up capacitors, 479
Spider, 331
Square root, 524
Square wave, 71, 73, 83, 97, 449, 466–467, 469
Squegging, 513
Stabilisation, 505, 513–516, 518
Stagger tuning, 371, 402
Standing wave, 266, 267
Standing-wave ratio, 267
Store, computer, 490, 498–500
Stray capacitance, 113, 124, 232, 234, 312, 346, 403, 457, 474, 476
Stub, 271
Subcarrier, 257, 440
Subscripts, 524
Substrate, 166, 464
Superheterodyne principle, 373–386
Superposition theorem, 451
Superscripts, 524
Surface-acoustic-wave filter, 399, 431
Surge impedance, 262
Susceptance, 95, 118
Sweep unit, 421
Swinging choke, 510
Switch, 28
  electronic, 483
Switch-mode power suppliers, 518
Symbols, algebraic, 520–524
  binary logic components, 494
  circuit diagram, 26, 28
  letter, 520–522, 523
  mathematical, 522
Sync. separator, 432–433
Synchronising circuits, 424, 432–435